Application of in-silico approaches to cardiovascular disease

Application of *in-silico* approaches to cardiovascular disease

Du, Jiangfeng
Application of *in-silico* approaches to cardiovascular disease by Jiangfeng Du, Maastricht:
Thesis Universiteit Maastricht, 2014
ISBN: 978-90-9028425-5.

Cover design: Jiangfeng Du
Printed by: Print & Copy Service, Sanmenxia, China

Application of *in-silico* approaches to cardiovascular disease

Dissertation

to obtain the Degree of Doctor at Maastricht University on the authority of Rector Magnificus, Prof. dr. L. L. G. Soete, in accordance with the decision of the Board of Deans of Maastricht University, to be defended in public in Maastricht on 8th of October 2014 at 10.00 a.m.

by

Jiangfeng Du

Born 10th of November 1982 in Sanmenxia, Henan, China

Promotor:

Prof. dr. Tilman M. Hackeng

Co- Promotor:

Dr. Gerry.A.F. Nicolaes

Assessment Committee:

Prof. dr. Erik Biessen (Chairman)

Prof. dr. Jan Glatz (Maastricht University)

Prof. dr. Siewert-Jan Marrink (University of Groningen)

Dr. Bruno Villoutreix (INSERM-Univ Paris, France)

Financial support by the Netherlands Heart Foundation for the publication of this thesis is gratefully acknowledged.

Contents

Chapter 1

General Introduction

1 Cardiovascular disease

Cardiovascular diseases (CVD) form a group of diseases that affect components of the cardiovascular system: the heart, blood or blood vessels. Examples of these diseases are atherosclerosis (1), hypertension (2), thrombosis (3) and haemophilia (4). CVD are the most common cause of death worldwide (5) and only in the United States there are an estimated number of 62 million people diagnosed with CVD (2). Records show that in 2008 around 17.3 million people died from CVD and it is estimated that 23 million people will die from CVD by 2030 annually worldwide (5, 6) according to the World Health Organization (WHO). Various risk factors for CVD exist, both genetically determined and acquired, such as age (7), air pollution (8), unhealthy diet (8-10), sex (11), lack of exercise (12), stress (13), genetic factors (14) and alcohol consumption (15). For normal homeostasis, blood coagulation is required and the coagulation factors, being proteins, should be properly expressed and functional otherwise it may lead to the development of thrombotic (3, 16) or bleeding disorders (4).

The cardiovascular system and immune system The cardiovascular system interacts with the immune system (17, 18). An Illustration for this interaction can be found not only in the fact that the two systems have evolved from a common ancestor system (19) but also in the observation that some diseases develop from the dysfunctioning of both systems (20). On one hand, certain pathogens or toll-like receptors are reported to be involved in the development of cardiovascular diseases by triggering a prothrombotic response or

by the induction of platelet aggregation (21, 22). On the other hand, coagulation processes are able to physically immobilize and catch invading pathogens in fibrin clots (23). Some coagulation factors are even directly involved in the innate immune system. For example, thrombin and factor X (FX) have pro-inflammatory properties as they induce the release of the several cytokines (24-27); the most notable case is that of the protease-activated receptors (PARs), which are able to mediate the inflammatory functions of a series of blood coagulation proteases, including thrombin (20, 28). Blood coagulation and host defense are thus intrinsically intertwined and both processes have been studied in the current work.

2 Introduction of *In silico* approaches

Besides the so-called *wetlab* experiments, *in silico* approaches are relatively new methods in the CVD field. *In silico* approaches are a common denominator for techniques and technologies that require computation and given the presence of silicium-based processors/computer chips, are referred to as "*in silico*". These techniques are very diverse and include methodologies applied for database buildings, virtual ligand screening (VLS) and drug design, 3D structure prediction and optimization, binding energy calculations and molecular dynamics (MD) and many more (Table 1.1). The following sections briefly introduce the *in silico* methods that have been applied in this dissertation.

Structural determinations

Biological functions of proteins are determined by their three-

dimensional (3D) structures (29-31). In order to perform their functions, proteins have to fold correctly into their specific conformations, and protein folding is driven by a number of non-covalent interaction forces including Van der Waals (VdW) forces, electrostatics forces, ionic interactions, hydrogen bonds, and hydrophobic packing (32-36).

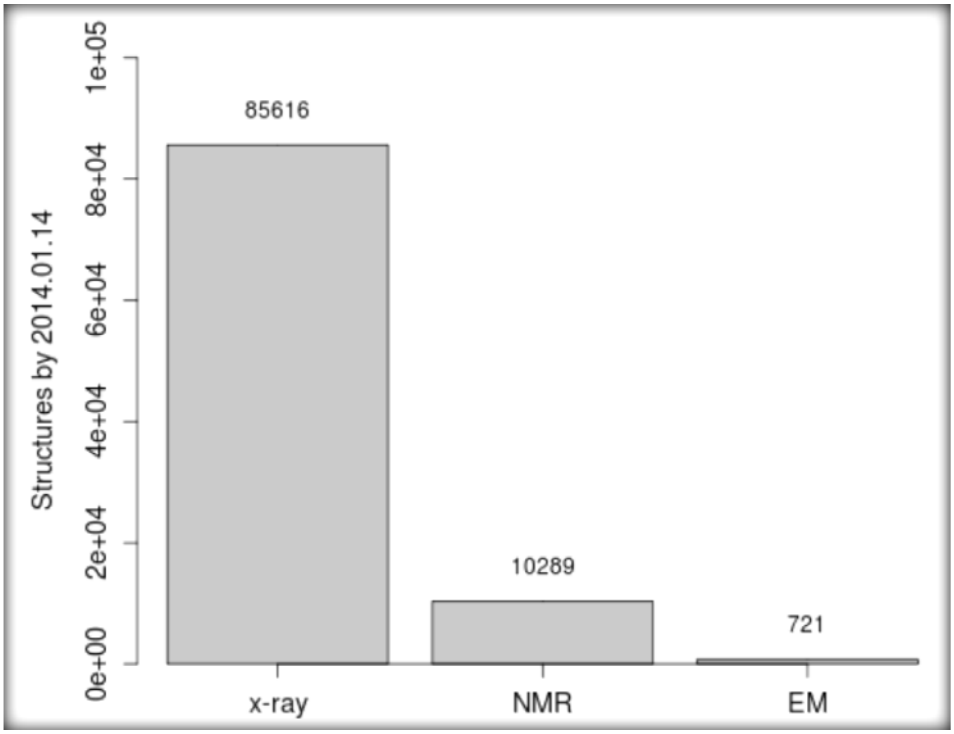


Figure 1.1 the numbers of 3D structures stored in RCSB protein data bank up to 14-01-2014. 88.6% of the total structures were identified by x-ray crystallography, 10.65% and 0.75% were identified by NMR and EM respectively.

Therefore, to rationally analyze protein functions, protein interactions, antigenic behavior, or the rational design of drugs (compounds)

against a protein target, a 3D structure of a target of interest is necessary. Experimental methods for structural determination are currently mainly X-ray crystallography, NMR spectroscopy and Cryo-electron microscope (EM).

The public repository of freely accessible protein structures, the Protein Data Bank or PDB (37) contains more than 96626 protein structures up to 14 January, 2014. Of these, around 88.6%, 10.65% and 0.75%, have been determined by X-ray crystallography, NMR spectroscopy and EM, respectively (Figure 1.1).

Homology modeling

Several factors limit the application of experimental methods to determine a protein's structure: 1, the methods are time consuming and costly, which is true for crystallography, NMR spectroscopy and Cryo-Electron Microscopy). 2, the structures of not all proteins can be studied by means of experimental structure determination, usually because of size- or stability limitations or a failure to find proper crystallization conditions. 3, failure to express proteins to sufficiently large amounts needed for structure determination. 4, the error in deciphering of experimental data into 3D model is usually inevitable, which urges new technologies to study proteins' structures. Besides, numerous proteins in nature require less time consuming techniques to determine the 3D structures.

Homology modeling is an *in-silico* approach that is able to generate 3D structure models of atomistic resolution for a desired target

(mostly a protein) based on the primary sequence (the amino acid sequence) and its relation to one or more homologous proteins of known 3D structures of (38). Thus, homology modeling is a supplementary method to obtain structural information that is useful in especially those cases when experimental methods fail. The concept of homology modeling is on the basis of two observations: 1, the primary sequence determines the secondary and tertiary protein structure (39). 2, between homologous proteins, the tertiary structure is more conserved than the primary sequence (40). A homology model can be considered trustworthy when the sequence identity between a given protein and its template fall in the “safe homology modeling zone” shown in Figure 1.2.

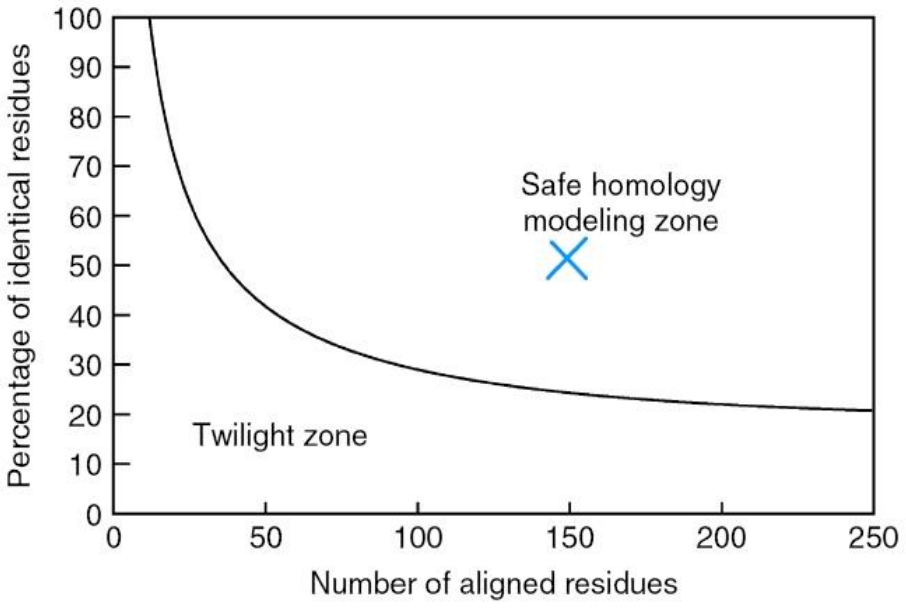


Figure 1.2 homology modeling quality control curve by the sequence identity and sequence length. The homology model is mostly trusted if the sequence alignment falls in the “Safe homology modeling zone” otherwise the template should not be used for homology modeling. The figure is taken from Hanka Venselaar, CMBI and the quality curve was derived from the work by Sander, C.

and R. Schneider in 1991(40, 41).

To construct a high quality homology model, one has to follow 6 procedures: **1 Template recognition.** To identify a proper template, a query sequence is aligned with each sequence of all the known structures in the PDB database (42) by a sequence alignment program, such as BLAST (43). Multiple sequence alignments may help to increase the correctness of the alignment between the query and the template sequence since extra information can be obtained from the other related sequences. **2 non-loop backbone generation.** The protein backbone structure of a model can be generated when the alignment information is present. When two aligned residues are identical, the 3D coordinates of the template residue including those of the side chain are copied from the template to the model's residue, while when two aligned residues differ, only the backbone coordinate is assigned to the aligned residue in the model. Notably, the 3D structure information from the PDB database may contain errors, and one has to be careful when deal with the template information. The PDB_redo database (44) aims to store protein 3D structure information in standardized formats, with improved R values and for optimized 3D structures, so it is advisable to obtain a template from PDB_redo database if possible. **3 loop modeling.** This is the most difficult part in homology modeling since loops are highly flexible, and can be classified into 56 groups according to their Ramachandran angles (45, 46). Two approaches are used to address loop structures: knowledge based loop prediction and energy based *ab-initio* loop prediction. Knowledge based loop prediction builds a loop model by similarity

search of known loop-conformations in the PDB database, followed by optimization of the model loop. Energy based *ab-initio* loop prediction builds a number of random loop conformations from scratch (*ab initio*) to sample the potential conformational space of a loop and then energy functions are used to evaluate the quality and likelihood of those loops. The best model can be further improved by methods like Monte Carlo energy minimization (47) and/or molecular dynamics (MD) simulations (48). **4 side chain modeling.** It is reported that similar residues share a similar C_α-C_β bond angle or even C_γ angle in a conserved region, so called “rotamer” (49, 50). Therefore, a side chain model can be built based on “rotamer” libraries and in most case this method is able to provide satisfactory results (51-53). **5 model optimization.** A correct side chain conformation prediction requires an accurate backbone conformation, which is in turn influenced by the side chain conformations. Therefore, one has to iteratively model the side chains and backbone conformations. For example, in chapter 4 and 5 we firstly built the backbone conformation of human IRAK-M death domain and then side chain conformations followed. In a second round, an optimized backbone conformation was built based on the side chain conformation, and then an improved side chain conformation was modeled based on an optimized backbone conformation. The procedure is performed iteratively until a satisfactory model, as estimated by model validation (see step 6) emerges. **6 model validation.** Several factors make the validation of a protein model structure a crucial procedure: a, the structure of the template may contain errors. b, the query sequence differs from the

template sequence which introduces uncertainty in model structure. c, rotamers of a residue mostly differ in their different backbone and local environments. d, the flexibility of loop regions require more validation rather than energy terms. e, the minimum state of an energy function does not always coincide with the best conformation of a given model. Many programs have been developed to evaluate a structure. These methods include various factors such as backbone packing quality, non-bonded interaction quality, 1D-3D profile consensus, atomic volume and bond angle/dihedral quality (Table 1.2) and can be used to judge the quality of a model structure.

Virtual ligand screening (VLS)

In drug discovery, high-throughput screening (HTS) has been successfully used to identify active compounds that are intended to interfere with the function of the protein target. Such molecules can be discovered after screening millions of compounds for a desired target. The screening efficiency of high throughput screening (HTS), which was introduced in the existing drug discovery pipelines in the 1980s, is relatively low as it required almost one year originally to screen 10,000 compounds against one target. With the development of the HTS, and automated laboratory methods, the screening time became faster and faster (54, 55) and in the 1990s, one week was enough to screen 10,000 compounds. In 21st century especially after 2010, a new design for HTS is able to increase the screening time over 1,000-fold. However, at the same time, more and more compounds are available from commercial vendors, with currently >35 million compounds available (Zinc database d.d. 17-3-

2014). With this large number of available molecules time considerations in HTS screening have become of lesser importance than the cost of screening itself and the selection of which compounds to include in the HTS campaign. Moreover, a specific HTS method needs to be developed for every new target. Thus to allow HTS to function optimally in the 21st century, including the new technologies currently available, complementary techniques have been developed.

Virtual ligand screening (VLS) is without doubt an ideal complementary technology to HTS. The *in-silico* approach is able to screen all compounds from a chemical compound database in order to obtain a pre-selection compounds which are most likely to bind to a target of interest (56-59). Two kinds of VLS approach in rational drug design can be distinguished: ligand-based VLS (LBVLS) and structure-based VLS (SBVLS) (60, 61).

Ligand-based VLS (LBVLS) If a list of ligands has been functionally studied (known active or inactive), the ligands' properties such as the molecular descriptors, 1D, 2D or 3D structural information can be used to select other compounds from an compound database (62, 63), which can be performed by various different algorithms such as similarity searching (64), pharmacophore mapping (65, 66) and 3D shape matching (67). Given a set of structural diverse ligands of a protein, it is possible to develop a quantitative structure activity relationship (QSAR) model, which can be used to define which parts of a model contribute to the function, and then identify even higher active compounds (68).

Structure-based VLS (SBVLS) SBVLS show its value when a 3D structure of the target protein is known (69), which enables small molecules to be docked into a defined binding pocket. The docked poses of every compound are evaluated by VLS scoring functions and the most likely ligands from the compiled list of docked and scored compounds can be selected for functional investigation. SBVLS can be further divided into two different docking methods: rigid and flexible docking methods. In rigid body docking, both ligand and receptor cannot change their conformations, which in turn require relatively little intensive computations (70, 71). In flexible docking, the compounds and/or the receptor are allowed to change their spatial shapes to fit their docking partners, which theoretically increases the docking accuracy but also increases the computational burden. Both docking approaches have their advantages and disadvantages and both docking approaches have generated some success stories (72-79).

Many software packages (Table 1.3), both commercial and non-commercial, are used for virtual ligand screening (VLS). The performances of docking programs, indispensable for performing any VLS campaign, have been compared in the literature (80) and each of these packages has its advantages and disadvantages (80). A general method that performs optimal on every target is currently not at hand. To approach the VLS docking campaign pragmatically, multi-step VLS protocols such as a combined FRED-Surflex docking procedure have been applied to obtain overall better results than with application of any of the docking tools individually (81). However, how to best combine both methods is a question that has not

sufficiently been addressed yet and which is systematically dealt with in Chapter 2.

Molecular dynamics (MD) simulation The level of maturity of a scientific field is reflected by the prominence of mathematics applied within that field. MD simulation is based on mathematics, physics and chemistry to study the properties of a molecular system by calculation of atomic interactions over time (by simple application of Newton's law or by a combination of quantum mechanics (QM) and molecular mechanics (MM)). It is only possible to conduct MD simulation after the target molecules properties, such as the composition, atom volume, mass and interaction parameters have been determined. If for all atoms of a given system the mass, position, and velocities are known, then MD simulation is able to calculate and simulate the motion of all atoms from the respective position coordinates and velocities for any atom at any time point during the simulation, resulting in the generation of a trajectory over time (82, 83). Historically, MD simulation can be tracked back to the 1950's, when Berni Alder published an article about MD simulation on a hard sphere system in J. Chem. Phys. It was not until 1974 that for the first time MD simulation was applied on the study of cell membrane pores and a solvent diffusion coefficient value could be calculated that was similar to an experimentally determined value (84). In 1977, the method was firstly applied to study a bovine pancreatic trypsin inhibitor published in Nature (85) and in the 1980s', the method was further applied to study ion transport across membranes (86), DNA molecules (87), water movement (88) and membrane-peptide interactions (89). The theoretical background of

the MD simulation is derived from the Newtonian equation for the motion of a particle:

$$E(\underline{R}) = \underbrace{\sum_{bonds} k_i^{bond}(r_i - r_0)^2}_{E_{bond}} + \underbrace{\sum_{angles} k_i^{angle}(\theta_i - \theta_0)^2}_{E_{angle}} + \underbrace{\sum_{dihedrals} k_i^{dihedral}(1 - \cos(n_i\phi_i + \xi_i))^2}_{E_{dihedral}} +$$

$$\underbrace{\sum_i \sum_{j \neq i} 4\xi_{ij}[(\frac{\sigma_{ij}}{r_{ij}})^{12} + (\frac{\sigma_{ij}}{r_{ij}})^6] + \sum_i \sum_{j \neq i} \frac{q_i q_j}{\xi r_j}}_{E_{non-bond}}$$

$E(bond)$ = the potential energy of a bond between two atoms i and j

$E(angle)$ = the potential energy of a bond angle between three atoms

$E(dihedral)$ = the potential energy regarding the torsional rotation between four atoms

$E(nonbond)$ = the potential energy between non - bond atoms i and j, refer to vdW and electrostatics

Assuming that the number of atoms in a system is N, there are 2* ((N-1)!) calculations for only non-bonded interactions. However, the non-bond energy drops to almost 0 if the distance of two atoms is beyond a value (e.g 1.2 nm), this threshold is used in many molecular dynamics force fields and described as “cutoff”. A proper cutoff is able to significantly decrease the calculation burden and does not impair the simulation accuracy.

Another time-consuming component of a MD calculation is the frequency to evaluate the force/energy, which refers to another important factor in simulation parameter setup: the time step. Theoretically, the smaller time step, the higher the simulation accuracy, but at a cost of more intensive calculation. Different MD simulation force fields may recommend different time steps for appropriate calculations, for example, for an atomistic MD simulation, the time step is generally around 1-3 femtosecond (fs) (90, 91).The calculation power of computer systems nowadays has been increasing enormously with the continuous development of computer

hardware, as well as though the advances in software. For example, the calculation time to simulate a system of 500 atoms for 9.8 picosecond (ps) in 1977 was 50,000 times longer than the same work done already in 2002 (82), and according to the Moore's law, 400,000 times longer than the calculation time in 2014. However, to study complicated biological systems such as proteins in the presence of lipid bilayers in an explicit solvent environment, protein-DNA complexes or enzyme complexes, it is not uncommon to simulate hundred thousand atoms or more simultaneously. Moreover, in most cases, the biological process of study requires longer timescales (e.g. microseconds or even milliseconds) rather than 10 ps. In fact, the gap of time scale and system size scale between biological processes (experimental measurements) and atomistic MD simulation remains one of the main hurdles to take in MD simulation.

A possible solution to decrease the gap is offered by a technique called coarse graining, which results in so-called "coarse-grained MD simulation". The coarse grained simulation applies a modified molecular dynamics force field in which the systems are represented by a reduced number of components and lesser number degrees of freedom. Due to its approximation of atomistic details, a coarse grained simulation requires less computational calculation expense as compared to atomistic simulations. The time steps in coarse grained simulation can range from 20-40 femtoseconds (fs) (92), which is another reason which makes that coarse grained simulations are less computationally expensive than atomistic simulations. As a result, a 1,000 fold improvement can be achieved in simulation time and system size when going from atomistically

detailed to coarse grained MD simulations. Usually, but depending on the research question, a coarse grained simulation expands both the simulation time from nanosecond scales to microsecond scale and the simulation system to hundreds thousands atoms. The approach enables the study of large complexes of biomolecules such as the interaction of G protein coupled receptors (GPCRs) with membranes (93-96), which are difficult to study by means of experimental/functional methods because of low yields of expression and practical difficulties in protein purification (97, 98). A non-exhaustive list of MD simulation packages has been presented here (Table 1.3).

3 Applications of *in-silico* approaches on cardiovascular diseases

As more and more proteins and other biomolecules and their 3D structures that are involved in the cardiovascular system are being identified, researchers now are interested in studying their structure-function relationships. Besides providing a rational explanation for the proper functioning of proteins, it is of interest to describe in atomistic details how these proteins and biomolecules interact with each other and then to translate such novel knowledge into novel drugs to interfere with protein function, as a potential means of therapy. While so-called *wetlab* experiments are in most cases able to provide useful insights into elucidation of structure-function relationships, this kind of experimentation usually gives indirect answers because of the complexity of the cardiovascular system. Financial considerations can form another drawback of more

traditional experimentation, for example, the generation, housing and analysis of transgenic animals, though extremely useful in many instances, is not always affordable for all laboratories. *In-silico* approaches such as homology modelling, virtual ligand screening, MD simulation have become popular tools that may assist in the study of cardiovascular diseases as a complementary to *wetlab* experiments (99). The number of publications that use *in-silico* approaches to study cardiovascular related targets is numerous, which makes it is impossible to present all of them in this section. Instead, several examples were chosen to illustrate the successful use of several methods. A recent study by Neculai and coworkers published in Nature (100) describes the structure of scavenger receptor class B type I (SR-BI) and CD36 by homology modeling approach, based on a crystal structure of lysosome membrane protein 2 (LIMP-2). SR-BI, CD36 and LIMP-2 are members of the CD36 superfamily, which is an important regulator of lipid metabolism and distinguish normal and modified lipoproteins, as well as pathogen-associated molecular pattern (PAMP) molecules. LIMP-2 contains 478 amino acids (101, 102) and the x-ray structure of LIMP-2 S36-I429 was identified, which is a beta-barrel motif with 15 helix and 17 strands. The sequence identity between human SR-BI and human LIMP-2 is 34% and the sequence identity between human CD36 and human LIMP-2 is 33%. The homology models prepared for SR-BI and CD36 feature two conserved disulfide bridges 1: (C312-C318 (LIMP-2), C321-C323 (SR-BI), C313-C322 (CD36); 2:C274-C329(LIMP-2), C280-C334(SR-BI), C272-C333(CD36)) that exist in all CD36 superfamily members. CD36 was proposed to

contain an additional disulfide bridge (C243-C311). Study of the homology models reveals clearly different electrostatic potential surfaces for the three members studied and may explain the respective biological functions. Other examples of the application of *in-silico* approaches can be found in the published works from our group (103), where 9 novel molecules were identified which are able to specifically inhibit the binding of the FVIII C2 domain to a model membrane by applying a virtual ligand screening approach and several other computational approaches such as 3D structure analysis, drug like pocket identification, small molecules optimization and ADMEt filter. In this type of research, invariably an x-ray structure of was optimized and further used to find pockets (for small molecules to bind) by using a consensus pocket-finding approach. Finally, one pocket was selected for virtual ligand screening and from a large database of chemical compounds an *in silico* selection was made by application of a multi-step docking protocol (81, 104). Final selection and optimization involves functional characterization and direct binding analysis and iterative rounds of compound optimization. This protocol has been successfully applied in our laboratory on a variety of protein targets: FV, FVIII, APC and TRAF6 (103, 105, 106).

4 Coagulation factor VIII (FVIII)

The coagulation pathway is one of the most conserved homeostatic systems throughout biology. Blood coagulation is aimed at the prevention of blood loss from damaged blood vessels through formation of blood clots, after injury to the blood vessel system.

Triggering of coagulation can occur through two independent processes: via the intrinsic- (involving Factor XII, XI, IX, VIII) and the extrinsic pathway (tissue factor, Factors VII) that culminate in a common pathway (involving fibrinogen, II, V, and X). These proteins, along with a series of anticoagulant proteins and inhibitors, work together in a carefully regulated and balanced set of biochemical reactions to maintain blood fluidity, but at the same time are capable of a rapid coagulant response after injury. Defects of any of these proteins can lead to bleeding tendencies (eg. Hemophilia A in case of FVIII deficiency) or excessive clotting (thrombosis, e.g. in case of deficiency of coagulation inhibitors).

Blood coagulation factor VIII (FVIII) is a large plasma glycoprotein (107) and its gene, the F8 gene, locates at the X chromosome, which indicates the FVIII relevant disorders are sex-linked diseases (108). The F8 gene is one of the largest genes and comprises 26 exons, which encode a signal peptide and the primary protein.

The protein is composed of 6 domains: A1-A2-B-A3-C1-C2 (from N to C terminus), containing 2332 amino acids. The A1, A2 and A3 domains share approximately 40% sequential identity to each other (109, 110). The large interconnecting B domain contains 908 amino acids, is heavily glycosylated and is removed upon activation of FVIII. The carboxyl terminal C domains are involved in binding to another plasma protein, von Willebrand factor (vWF) and in binding to negatively charged membrane surfaces (111-115). Despite the similar nomenclature, the FVIII C domains are distinct from the C domains of signal-transduction proteins such as protein kinase C (116). Much

unlike other well-described membrane binding domains involved in blood coagulation, the FVIII C domains are not vitamin K-dependent domains, and C domains do not require Ca^{2+} to express their membrane binding properties (117). Eight disulfide bonds contribute the structural integrity, with two disulfide bridges located in the A1, the A2 and in the A3 domain and one located in each of the two C domains (118).

Factor VIII is expressed by several different tissues including the spleen, the lymph nodes, the liver and kidney (119, 120), of which the liver and kidney are regarded as the primary sources of the FVIII production (121). Physiologically, non-activated FVIII circulates in a complex with vWF in the blood (122, 123), which prevents FVIII from being cleaved and from being taken out of circulation (109).

In the presence of thrombin, several proteolytic cleavages occur at specific sites of FVIII (Figure 1.4): R372, R740, and R1689. Dissociation of vWF and FVIII occurs when R1689 is cleaved. The B domain is released when cleavage occurs at R740. A further proteolytic cleavage at R372 allows the A1 domain to separate from the A2 domain. Together these events result in the activation of FVIII to FVIIIa (124). FVIIIa serves as the non-enzymatic cofactor to the serine enzyme FIXa in the so-called intrinsic tenase complex. The tenase complex activates FX into FXa, one of the essential steps that ultimately lead to effective clot formation.

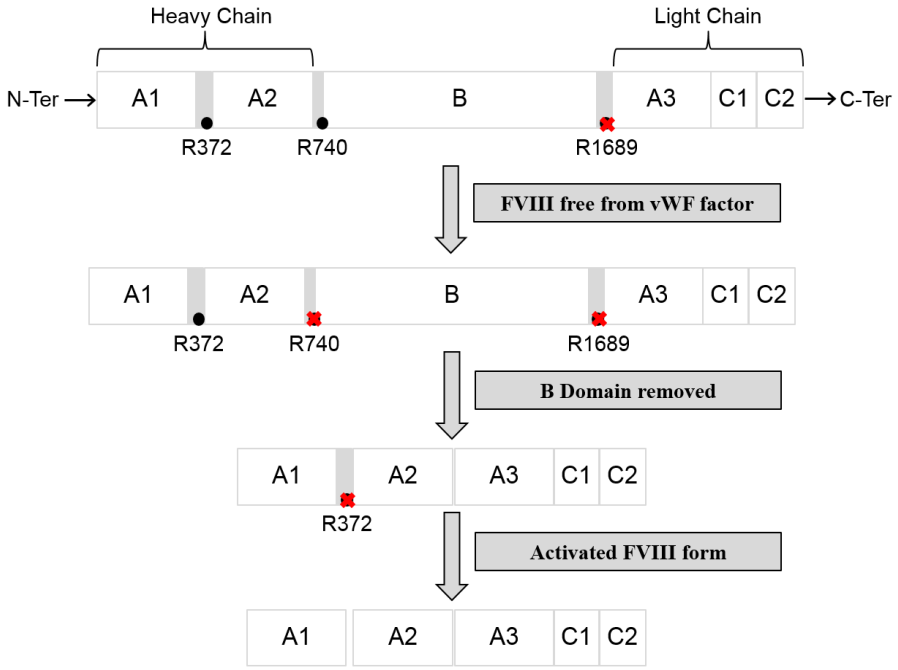


Figure 1.3 Primary domain organization of FVIII during proteolysis. The domain sizes are indicated by their box lengths. A1 and A2 form FVIII heavy chain and A3, C1 and C2 form the light chain. When thrombin presence, several proteolytic cleavages occur on specific sites of FVIII: R372, R740, and R1689. Dissociation of vWF and FVIII occurs after the cleavage of R1689. B chain is released when R740 is cleaved. Cleavage of R372 allows the A1 domain to separate from the A2 domain, which results in the activation of FVIII into FVIIIa (125).

After activation, when there is no vWF to stabilize FVIII and in the absence of covalent bonding between the A1 and A2 domains, the FVIII activity is rapidly lost , a phenomenon which is functionally favorable in order to avoid over-activation of FXa (126). Moreover, to regulate FVIIIa activity, FVIIIa is proteolytically inactivated by activated protein C (APC), FIXa, FXa (124, 127, 128).

For the expression of its cofactor activity, FVIIIa needs to bind to a

negatively charged membrane surface, and upon binding, the conversion of Factor X to factor Xa is enhanced over 100,000-fold (129). The C1 and C2 domains of FVIII are the main membrane-binding regions of the cofactor, of which C2 binding has been hypothesized as most important (113-115, 130-134).

5 IRAK family

Immunology and haemostasis are distinct but very much intertwined processes. A well-described example comes from the observation that severe sepsis, a disease characterized by a systemic inflammation, is associated with disseminated intravascular coagulation (DIC) which can result in a life-threatening coagulopathy and in organ failure (135). In particular the recent findings that the anticoagulant protein APC and the procoagulant proteins thrombin and FXa have multiple biological effects in haemostasis as well as in inflammatory disease has led to the realization that what used to be regarded as different disease entities (CVD and inflammation) may reflect different sides of a same coin. Likewise, the pathogenesis of atherosclerotic disease has clear components from both haemostatic and immunological origin (136-138).

The interleukin-1 receptor-associated kinase (IRAK) family belongs to the protein family of the protein kinases (139, 140). IRAK proteins serve as signal transducers for interleukin-1 (IL-1) and are involved in signaling innate immune responses from Toll-like receptors (TLR), a type of pattern recognition receptors. The IRAK family has four different members: IRAK-1, IRAK-2, IRAK-3 (IRAK-M) and IRAK-4.

IRAK-1 was first described by Croston and Cao in 1995 (141), who found that the IL-I receptor cannot activate NF-kappa B in the absence of IRAK-1. One year later, an IRAK-1 paper from the same group was published in Science (142), where it is described that IRAK is a kinase associated with IL-1R, sharing similarity with Pelle, another protein kinase in Drosophila. Two years later, IRAK-2 has been identified by an expressed sequence tag (EST) database analysis (143). Four years later, another IRAK member was identified, the tissue specific IRAK-3, also called IRAK-M, since it mainly presents in cells of monomyeloic origin. In 2002, almost 7 years after the first IRAK member being identified, Suzuki and his coworkers discovered a novel IRAK protein (IRAK-4) by analyzing gene targeting studies and further concluded the function of IRAK-4 is to phosphorylate IRAK-1 and to regulate signal transduction (144). IRAK-1 and IRAK-4 contain active kinase subunits whereas IRAK-2 and IRAK-M remarkably contain inactive kinase subunits.

All IRAK members mediate activation of nuclear factor-kB (NF-kB) and mitogen-activated protein kinase (MAPK) via the TLR/IL-1 pathway (145). The signal transduction pathways initiated by toll/IL-1 receptor family members ultimately lead to the activation of transcription factors such as activator protein 1 (AP-1) or NF-kB, which contribute to the establishment of an immune response and also regulate target genes such as HO-1, COX-2, ecSOD, iNOS whose products are involved into myocardial protection (146).

In the TLR pathway, IRAK family members interact with TLR adapter proteins such as myeloid differentiation primary response protein

(MyD88), and also with tumor necrosis factor receptor associated factor 6 (TRAF6) (Figure 1.3). Both IRAK-1 and IRAK-2 can bind to the Toll-like receptor (IL-1R) and trigger intracellular signaling cascades that lead to transcriptional up-regulation and mRNA stabilization. The formation of the IRAK-IL-1R complex requires the presence of IL-1RAcP and MyD88 (147, 148). IRAK-4 forms a complex with IRAK1 and the IL-1 receptor to improve the efficient recruitment of IRAK-1 and while doing so it also phosphorylates the IRAK-1 protein. IRAK-M is thought to inhibit the dissociation of IRAK-1 and IRAK-4 from the Toll-like receptor signaling complex by either inhibiting the phosphorylation of IRAK-1 and IRAK-4 or by stabilization of the receptor complex. However, it appears as if IRAK-M has more functions than only the inhibition of the formation of IRAKs-MyD88 complexes. For example, IRAK-M can form a unique complex with IRAK-4 and mediate the MEKK3 activation pathway (149). IRAK-M is then able to stabilize MKP-1 and further down regulate the non-canonical NF-kappa B pathway (150). It has been described that a lack of IRAK-M leads to vulnerability to bacterial pneumonia (151-153). There are many observations listed in literature that concern the involvement of IRAK related proteins in the cardiovascular system. To name only a few: inhibition of IRAK-1 protects the mouse and human small intestine against ischemia/reperfusion injury (154), TLR9 inhibition can protect the liver from ischemia-reperfusion injury (155), TLR4 is responsible for protection against intestinal ischemia reperfusion (156), down regulation of IRAK4 and NF-kappa B level may protect against hepatic ischemia-reperfusion injury (157). A general scheme (Figure

1.3) illustrates the functions of IRAK family members in the NF-κB pathway and further refers to myocardial protection.

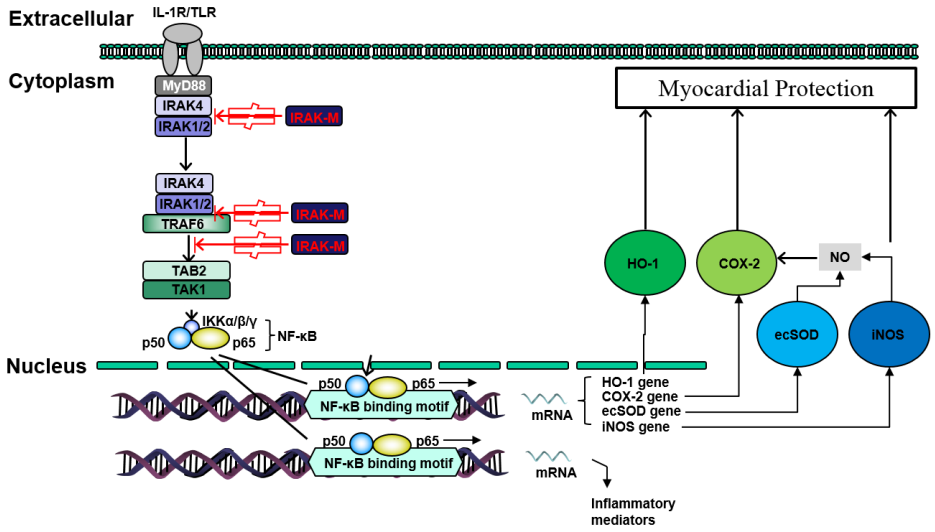


Figure 1.4 Schematic representation of the NF-κB pathway and myocardial protection. Stimulation of IL-1R/TLR triggers the association of myeloid differentiation primary-response protein 88 (MyD88), which in turn recruits IRAK4 (IL-1R-associated kinase 4). IRAK1 and IRAK2 from complex with IRAK4. The complex of IRAK1, 2 and 4 then leaves from MyD88 after the phosphorylation and further form a new complex with tumor-necrosis-factor-receptor-associated factor 6 (TRAF6), which phosphorylates transforming-growth-factor-β-activated kinase (TAK1) and TAK1-binding protein 1 (TAB2), which further phosphorylates the inhibitor of nuclear factor-κB (IκB)-kinase complex (IKK-α, IKK-β and IKK-γ) complex, which then phosphorylate inhibitor of κB (IκB) and then destroy IκB. It allows NF-κB to translocate to the nucleus and induce the expression of its target genes, which result in transcriptional activation of *iNOS, COX-2*, *HO-1*, and *ecSOD*, leading to the synthesis of the respective proteins, which further protects the myocardium. IRAK-M is able to negatively regulate the NF-κB pathway by either prohibiting the formation of the complex of IRAK-M and IRAK1/2 or prohibiting the association of the complex of IRAK4 and MyD88. Directly binding to TRAF6 by IRAK-M suppresses TRAF6's phosphorylation.

Table 1.1 Major research areas by *in-silico* approaches

Major Field	Successful example
Sequence analysis	Human genome project (158, 159); BLAST (43); haemophilus influenzae genome (160)
Genome annotation	Protein-coding genes searching (161, 162)
Evolutionary biology	Evolutionary models (163)
Literature analysis	Pubmed (164); Protein Data Bank (42); PDB_redo (44)
Gene expression	Post process of the gene expression data
Analysis of regulation	promoter analysis (165, 166)
Analysis of protein expression	Analysis of protein microarray and mass spectrometry data (167)
Mutagenesis	SNP detector (168) ; Mutation prediction (169)
Comparative genomics	orthology analysis (170)
system biology	The simulation of cellular processes ((171, 172)
High-throughput	flow cytometry analysis (173)
Structural bioinformatics	Homology modelling (45); Virtual ligand screening (59, 174, 175); Molecular dynamics (82)

Table1.2 programs for protein structure evaluation

Validation Tools	Validation of
PROCHECK	backbone quality and Ramachandran value (176)
ERRAT	Non-bonded interaction quality (177)
Verify3D; 3DCA	1D-3D structure profiles (178-180)
PROVE	Atomic volume check (181)
PROSESS	Packing quality, torsion angle, H-bonds (182)
WHAT CHECK	Packing quality check (183)
PDBREPORT	Packing quality check and database (183)
SFCHECK;EDS	R-value related check and database (184, 185)
ProQ	A neural network predictor (186)
ANOLEA	Atomic non-local mean force potential assessment (187)
VADAR	Volume, area, dihedral angle (188)
PDBsum	A combination of EDS and PDBREPORT (189)

Table 1.3 a non-exhaustive list of commonly used programs, software and packages for molecular mechanics calculations.

Package name	Main Function	Comments	License and References
ICM	HM, View, Min, PPI, PF,	Multifunction packages	Commercial (190)

	VLS		
MOE	HM, View, Min, MD, VLS, QSAR	Multifunction packages	Commercial (191)
YASARA	HM, View, Min, MD	Based on what-if software	Commercial (192)
AMBER	Min, MD	A set of force fields and a MD simulation package	Commercial (193)
CHARMM	Min, MD	Accelrys company	Commercial (194)
GROMACS	Min, MD	Fast, stable and variety force fields including coarse-grained simulation	Free (90, 195)
CPHmodels	HM	Profile-profile alignment guided by secondary structure and exposure predictions.	Online server (196)
ChunkTASSER	HM	An ab-initio predictor	Online server (197)
ROBETTA	HM	Both ab-initio and comparative model predictor for small protein	Online server (198, 199)
ModBase	HM	A database of comparative models	Free (200)
HHpred	HM	Hidden Markov Models based predictor	Online server (201)
phyre	HM	Profile-profile matching algorithm	Online server (202)
SWISS-MODEL	HM	Hidden Markov Models based predictor	Online server (203)
Modeller	HM	Spatial restraint, optimization, multiple alignment, de novo	Free (48, 204)
ESyPred3D	HM	Neural network sequence alignment	Online server (205)
SCWRL	HM	Kernel density estimate rotamer library; hydrogen bond function; soft vdW potential;	Online server (206)
PyMOL	View	The most popular visualization program with lots of plugin	Commercial (207)
VMD	View	a long trajectory can be visualized	Free (208)
COLORADO-3D	View	Online visualization tools for structural features (combined with other programs)	http://asia2.genesilico.pl/colorado3d/
ODA	PPI	Solvent accessibility	Commercial

			(190)
Cons-PPISP	PPI	Neural network	Online server (209)
Struct2Net	PPI	Primary sequence based	Online server (210)
PPI-Pred	PPI	Support vector machine algorithm	Online server (211)
FRED	SBVLS	Rigid; Fast	License needed (212)
Surflex	SBVLS	Flexible for ligand	License needed (213)
GOLD	SBVLS	Flexible for ligand and receptor	Commercial (214)
DOCK	SBVLS	Flexible for ligand and receptor	License needed (215)
Glide	SBVLS	Flexible for ligand and receptor	Commercial (216)
eHiTS	SBVLS	Flexible for ligand	Commercial (217)
HYBRID	SBVLS	Fast; ligand-guided	License needed (218)
LigandFit	SBVLS	CHARMM; Monte-Carlo; Rigid	Commercial (219)
Pharmer	LBVLS	Pharmacophore search; fast	Free (220)
FITTED	SBVLS	Genetic algorithm; Flexible for ligand and receptor	Free (221)
FlexX	SBVLS	Flexible for ligand; partial flexible for receptor	License needed (222)
Vina	SBVLS	Monte-Carlo; Flexible for ligand and side chains	License needed (223)
PocketFinder	PF	Shape and physicochemical; Implemented with MolSoft	Commercial (224)
Qsitefinder	PF	Energy function with VdW probe	Online server (225)
MetaPocket	PF	Consensus of LIGSITE, PASS, Q-SiteFinder, and SURFNET	Online server (226)
DogSiteScorer	PF	Geometric and physicochemical matches with Gaussian function	Online server (227)

Notes: HM: homology modelling; View: molecular visualization tool; Min: energy minimization; MD: molecular dynamics simulation; PPI: protein-protein interaction prediction; PF: druggability pocket finder; LBVLS: ligand based virtual ligand screening; SBVLS: structure based virtual ligand screening; DB: database.

References

1. Ross R (1993) The pathogenesis of atherosclerosis: a perspective for the 1990s. *Nature* 362(6423):801-809.
2. Nabel EG (2003) Cardiovascular disease. *The New England journal of medicine* 349(1):60-72.
3. Furie B & Furie BC (2008) Mechanisms of thrombus formation. *The New England journal of medicine* 359(9):938-949.
4. Bowen DJ (2002) Haemophilia A and haemophilia B: molecular insights. *Molecular pathology : MP* 55(2):127-144.
5. Geneva (2011) Global status report on noncommunicable diseases 2010. *World Health Organization*
6. Mathers CD LD (2006) Projections of global mortality and burden of disease from 2002 to 2030. *PLoS Med* 3.
7. Mackay M, Mendis (2004) The Atlas of Heart Disease and Stroke. *World Health Organization*.
8. Ignarro LJ, Balestrieri ML, & Napoli C (2007) Nutrition, physical activity, and cardiovascular disease: an update. *Cardiovascular research* 73(2):326-340.
9. Walker C & Reamy BV (2009) Diets for cardiovascular disease prevention: what is the evidence? *American family physician* 79(7):571-578.
10. Nordmann AJ, *et al.* (2011) Meta-analysis comparing Mediterranean to low-fat diets for modification of cardiovascular risk factors. *The American journal of medicine* 124(9):841-851 e842.
11. Finegold JA, Asaria P, & Francis DP (2013) Mortality from ischaemic heart disease by country, region, and age: statistics from World Health Organisation and United Nations. *International journal of cardiology* 168(2):934-945.
12. Buttar HS, Li T, & Ravi N (2005) Prevention of cardiovascular diseases: Role of exercise, dietary interventions, obesity and smoking cessation. *Experimental and clinical cardiology* 10(4):229-249.
13. Dimsdale JE (2008) Psychological stress and cardiovascular disease. *Journal of the American College of Cardiology* 51(13):1237-1246.
14. Cambien F & Tiret L (2007) Genetics of cardiovascular diseases: from single mutations to the whole genome. *Circulation* 116(15):1714-1724.
15. Klatsky AL (2009) Alcohol and cardiovascular diseases. *Expert review of cardiovascular therapy* 7(5):499-506.
16. Furie B & Furie BC (2005) Thrombus formation in vivo. *The Journal of clinical investigation* 115(12):3355-3362.
17. Opal SM & Esmon CT (2003) Bench-to-bedside review: functional relationships between coagulation and the innate immune response and their respective roles in the pathogenesis of sepsis. *Critical care* 7(1):23-38.
18. Potaczek DP (2014) Links between allergy and cardiovascular or hemostatic system. *International journal of cardiology* 170(3):278-285.
19. Krem MM & Di Cera E (2002) Evolution of enzyme cascades from embryonic development to blood coagulation. *Trends in biochemical sciences* 27(2):67-74.
20. Demetz G & Ott I (2012) The Interface between Inflammation and Coagulation in Cardiovascular Disease. *International journal of inflammation* 2012:860301.
21. Herzberg MC (2001) Coagulation and thrombosis in cardiovascular disease: plausible contributions of infectious agents. *Annals of periodontology / the*

American Academy of Periodontology 6(1):16-19.
22. Cook DN, Pisetsky DS, & Schwartz DA (2004) Toll-like receptors in the pathogenesis of human disease. *Nature immunology* 5(10):975-979.
23. Iwanaga S (2002) The molecular basis of innate immunity in the horseshoe crab. *Current opinion in immunology* 14(1):87-95.
24. Drake WT, Lopes NN, Fenton JW, 2nd, & Issekutz AC (1992) Thrombin enhancement of interleukin-1 and tumor necrosis factor-alpha induced polymorphonuclear leukocyte migration. *Laboratory investigation; a journal of technical methods and pathology* 67(5):617-627.
25. Fujita T, *et al.* (2008) Thrombin enhances the production of monocyte chemoattractant protein-1 and macrophage inflammatory protein-2 in cultured rat glomerular epithelial cells. *Nephrology, dialysis, transplantation : official publication of the European Dialysis and Transplant Association - European Renal Association* 23(11):3412-3417.
26. Wadgaonkar R, Somnay K, & Garcia JG (2008) Thrombin induced secretion of macrophage migration inhibitory factor (MIF) and its effect on nuclear signaling in endothelium. *Journal of cellular biochemistry* 105(5):1279-1288.
27. Krupiczojc MA, Scotton CJ, & Chambers RC (2008) Coagulation signalling following tissue injury: focus on the role of factor Xa. *The international journal of biochemistry & cell biology* 40(6-7):1228-1237.
28. Shpacovitch V, Feld M, Hollenberg MD, Luger TA, & Steinhoff M (2008) Role of protease-activated receptors in inflammatory responses, innate and adaptive immunity. *Journal of leukocyte biology* 83(6):1309-1322.
29. Lee D, Redfern O, & Orengo C (2007) Predicting protein function from sequence and structure. *Nature reviews. Molecular cell biology* 8(12):995-1005.
30. Watson JD, Laskowski RA, & Thornton JM (2005) Predicting protein function from sequence and structural data. *Current opinion in structural biology* 15(3):275-284.
31. Fu H, Subramanian RR, & Masters SC (2000) 14-3-3 proteins: structure, function, and regulation. *Annual review of pharmacology and toxicology* 40:617-647.
32. Anfinsen CB (1972) The formation and stabilization of protein structure. *Biochem J* 128(4):737-749.
33. Selkoe DJ (2003) Folding proteins in fatal ways. *Nature* 426(6968):900-904.
34. Anfinsen CB (1973) Principles that govern the folding of protein chains. *Science* 181(4096):223-230.
35. Pace CN, Shirley BA, McNutt M, & Gajiwala K (1996) Forces contributing to the conformational stability of proteins. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* 10(1):75-83.
36. Rose GD, Fleming PJ, Banavar JR, & Maritan A (2006) A backbone-based theory of protein folding. *Proceedings of the National Academy of Sciences of the United States of America* 103(45):16623-16633.
37. Bernstein FC, *et al.* (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. *Journal of molecular biology* 112(3):535-542.
38. Chothia C & Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. *The EMBO journal* 5(4):823-826.
39. Hirs CH, Moore S, & Stein WH (1960) The sequence of the amino acid residues in performic acid-oxidized ribonuclease. *J Biol Chem* 235:633-647.
40. Sander C & Schneider R (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* 9(1):56-68.

41. Rost B (1999) Twilight zone of protein sequence alignments. *Protein engineering* 12(2):85-94.
42. Bernstein FC, *et al.* (1977) The Protein Data Bank. A computer-based archival file for macromolecular structures. *European journal of biochemistry / FEBS* 80(2):319-324.
43. Altschul SF, Gish W, Miller W, Myers EW, & Lipman DJ (1990) Basic local alignment search tool. *Journal of molecular biology* 215(3):403-410.
44. Joosten RP, Joosten K, Murshudov GN, & Perrakis A (2012) PDB_REDO: constructive validation, more than just looking for errors. *Acta crystallographica. Section D, Biological crystallography* 68(Pt 4):484-496.
45. Rodriguez R, Chinea G, Lopez N, Pons T, & Vriend G (1998) Homology modeling, model and software evaluation: three related resources. *Bioinformatics* 14(6):523-528.
46. Oliva B, Bates PA, Querol E, Aviles FX, & Sternberg MJ (1997) An automated classification of the structure of protein loops. *Journal of molecular biology* 266(4):814-830.
47. Simons KT, Bonneau R, Ruczinski I, & Baker D (1999) Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins* Suppl 3:171-176.
48. Fiser A, Do RK, & Sali A (2000) Modeling of loops in protein structures. *Protein Sci* 9(9):1753-1773.
49. Summers NL, Carlson WD, & Karplus M (1987) Analysis of side-chain orientations in homologous proteins. *Journal of molecular biology* 196(1):175-198.
50. Lee C & Subbiah S (1991) Prediction of protein side-chain conformation by packing optimization. *Journal of molecular biology* 217(2):373-388.
51. Liang S & Grishin NV (2002) Side-chain modeling with an optimized scoring function. *Protein Sci* 11(2):322-331.
52. Xiang Z (2006) Advances in homology protein structure modeling. *Current protein & peptide science* 7(3):217-227.
53. Bower MJ, Cohen FE, & Dunbrack RL, Jr. (1997) Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. *Journal of molecular biology* 267(5):1268-1282.
54. Major J (1998) Challenges and Opportunities in High Throughput Screening: Implications for New Technologies. *journal of biomolecular screening* 3:13-17.
55. Agresti JJ, *et al.* (2010) Ultrahigh-throughput screening in drop-based microfluidics for directed evolution. *Proceedings of the National Academy of Sciences of the United States of America* 107(9):4004-4009.
56. Rester U (2008) From virtuality to reality - Virtual screening in lead discovery and lead optimization: a medicinal chemistry perspective. *Current opinion in drug discovery & development* 11(4):559-568.
57. Rollinger JM, Stuppner H, & Langer T (2008) Virtual screening for the discovery of bioactive natural products. *Progress in drug research. Fortschritte der Arzneimittelforschung. Progres des recherches pharmaceutiques* 65:211, 213-249.
58. Klebe G (2006) Virtual ligand screening: strategies, perspectives and limitations. *Drug discovery today* 11(13-14):580-594.
59. Schneider G (2010) Virtual screening: an endless staircase? *Nature reviews. Drug discovery* 9(4):273-276.
60. McInnes C (2007) Virtual screening strategies in drug discovery. *Current opinion in chemical biology* 11(5):494-502.
61. Cavasotto CN & Orry AJ (2007) Ligand docking and structure-based virtual

screening in drug discovery. *Current topics in medicinal chemistry* 7(10):1006-1014.
62. Reddy AS, Pati SP, Kumar PP, Pradeep HN, & Sastry GN (2007) Virtual screening in drug discovery -- a computational perspective. *Current protein & peptide science* 8(4):329-351.
63. Lengauer T, Lemmen C, Rarey M, & Zimmermann M (2004) Novel technologies for virtual screening. *Drug discovery today* 9(1):27-34.
64. Quintus F, Sperandio O, Grynberg J, Petitjean M, & Tuffery P (2009) Ligand scaffold hopping combining 3D maximal substructure search and molecular similarity. *BMC bioinformatics* 10:245.
65. Sun H (2008) Pharmacophore-based virtual screening. *Current medicinal chemistry* 15(10):1018-1024.
66. Mason JS, Good AC, & Martin EJ (2001) 3-D pharmacophores in drug discovery. *Current pharmaceutical design* 7(7):567-597.
67. Srinivasan J, *et al.* (2002) Evaluation of a novel shape-based computational filter for lead evolution: application to thrombin inhibitors. *Journal of medicinal chemistry* 45(12):2494-2500.
68. Verma J, Khedkar VM, & Coutinho EC (2010) 3D-QSAR in drug design--a review. *Current topics in medicinal chemistry* 10(1):95-115.
69. Villoutreix BO, Eudes R, & Miteva MA (2009) Structure-based virtual ligand screening: recent success stories. *Combinatorial chemistry & high throughput screening* 12(10):1000-1016.
70. Perola E & Charifson PS (2004) Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding. *Journal of medicinal chemistry* 47(10):2499-2510.
71. Giganti D, *et al.* (2010) Comparative evaluation of 3D virtual ligand screening methods: impact of the molecular alignment on enrichment. *Journal of chemical information and modeling* 50(6):992-1004.
72. McNally VA, *et al.* (2003) Identification of a novel class of inhibitor of human and Escherichia coli thymidine phosphorylase by in silico screening. *Bioorganic & medicinal chemistry letters* 13(21):3705-3709.
73. Wang J, Kang X, Kuntz ID, & Kollman PA (2005) Hierarchical database screenings for HIV-1 reverse transcriptase using a pharmacophore model, rigid docking, solvation docking, and MM-PB/SA. *Journal of medicinal chemistry* 48(7):2432-2444.
74. Pan H, Agarwalla S, Moustakas DT, Finer-Moore J, & Stroud RM (2003) Structure of tRNA pseudouridine synthase TruB and its RNA complex: RNA recognition through a combination of rigid docking and induced fit. *Proceedings of the National Academy of Sciences of the United States of America* 100(22):12648-12653.
75. Sauton N, Lagorce D, Villoutreix BO, & Miteva MA (2008) MS-DOCK: accurate multiple conformation generator and rigid docking protocol for multi-step virtual ligand screening. *BMC bioinformatics* 9:184.
76. Mysinger MM, *et al.* (2012) Structure-based ligand discovery for the protein-protein interface of chemokine receptor CXCR4. *Proceedings of the National Academy of Sciences of the United States of America* 109(14):5517-5522.
77. Gehlhaar DK, *et al.* (1995) Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming. *Chemistry & biology* 2(5):317-324.

78. Cerqueira NM, Bras NF, Fernandes PA, & Ramos MJ (2009) MADAMM: a multistaged docking with an automated molecular modeling protocol. *Proteins* 74(1):192-206.
79. Dey R & Chen L (2011) In search of allosteric modulators of a7-nAChR by solvent density guided virtual screening. *Journal of biomolecular structure & dynamics* 28(5):695-715.
80. Zhou Z, Felts AK, Friesner RA, & Levy RM (2007) Comparative performance of several flexible docking programs and scoring functions: enrichment studies for a diverse set of pharmaceutically relevant targets. *Journal of chemical information and modeling* 47(4):1599-1608.
81. Miteva MA, Lee WH, Montes MO, & Villoutreix BO (2005) Fast structure-based virtual ligand screening combining FRED, DOCK, and Surflex. *Journal of medicinal chemistry* 48(19):6012-6022.
82. Karplus M & McCammon JA (2002) Molecular dynamics simulations of biomolecules. *Nature structural biology* 9(9):646-652.
83. Borhani DW & Shaw DE (2012) The future of molecular dynamics simulations in drug discovery. *Journal of computer-aided molecular design* 26(1):15-26.
84. Levitt DG & Subramanian G (1974) A new theory of transport for cell membrane pores. II. Exact results and computer simulation (molecular dynamics). *Biochim Biophys Acta* 373(1):132-140.
85. McCammon JA, Gelin BR, & Karplus M (1977) Dynamics of folded proteins. *Nature* 267(5612):585-590.
86. Fischer W, Brickmann J, & Lauger P (1981) Molecular dynamics study of ion transport in transmembrane protein channels. *Biophysical chemistry* 13(2):105-116.
87. Levitt M (1983) Computer simulation of DNA double-helix dynamics. *Cold Spring Harbor symposia on quantitative biology* 47 Pt 1:251-262.
88. Aityan SK & Chizmadzhev Yu A (1986) Simulation of molecular dynamics of water movement in ion channels. *General physiology and biophysics* 5(3):213-229.
89. Edholm O & Johansson J (1987) Lipid bilayer polypeptide interactions studied by molecular dynamics simulation. *European biophysics journal : EBJ* 14(4):203-209.
90. Van Der Spoel D, *et al.* (2005) GROMACS: fast, flexible, and free. *Journal of computational chemistry* 26(16):1701-1718.
91. Darian E & Gannett PM (2005) Application of molecular dynamics simulations to spin-labeled oligonucleotides. *Journal of biomolecular structure & dynamics* 22(5):579-593.
92. Marrink SJ, Risselada HJ, Yefimov S, Tieleman DP, & de Vries AH (2007) The MARTINI force field: coarse grained model for biomolecular simulations. *J Phys Chem B* 111(27):7812-7824.
93. Simpson LM, Taddese B, Wall ID, & Reynolds CA (2010) Bioinformatics and molecular modelling approaches to GPCR oligomerization. *Current opinion in pharmacology* 10(1):30-37.
94. Periole X, Knepp AM, Sakmar TP, Marrink SJ, & Huber T (2012) Structural determinants of the supramolecular organization of G protein-coupled receptors in bilayers. *Journal of the American Chemical Society* 134(26):10959-10965.
95. Mondal S, *et al.* (2013) Membrane driven spatial organization of GPCRs. *Scientific reports* 3:2909.
96. Ghosh A, Sonavane U, & Joshi R (2013) Multiscale modelling to understand the self-assembly mechanism of human beta2-adrenergic receptor in lipid bilayer. *Computational biology and chemistry* 48C:29-39.

97. Structural Genomics Consortium AeFdMBea (2008) Protein production and purification. *Nature methods* 5:135-146.
98. Wagner S, Bader ML, Drew D, & de Gier JW (2006) Rationalizing membrane protein overexpression. *Trends in biotechnology* 24(8):364-371.
99. Ataullakhanov FI & Panteleev MA (2005) Mathematical modeling and computer simulation in blood coagulation. *Pathophysiology of haemostasis and thrombosis* 34(2-3):60-70.
100. Neculai D, *et al.* (2013) Structure of LIMP-2 provides functional insights with implications for SR-BI and CD36. *Nature* 504(7478):172-176.
101. Fujita H, *et al.* (1992) Isolation and sequencing of a cDNA clone encoding the 85 kDa human lysosomal sialoglycoprotein (hLGP85) in human metastatic pancreas islet tumor cells. *Biochemical and biophysical research communications* 184(2):604-611.
102. Calvo D, Dopazo J, & Vega MA (1995) The CD36, CLA-1 (CD36L1), and LIMPII (CD36L2) gene family: cellular distribution, chromosomal location, and genetic evolution. *Genomics* 25(1):100-106.
103. Nicolaes GA, *et al.* (2014) Rational design of small molecules targeting the C2 domain of coagulation factor VIII. *Blood* 123(1):113-120.
104. Du J, Bleylevens IW, Bitorina AV, Wichapong K, & Nicolaes GA (2014) Optimization of compound ranking for structure-based virtual ligand screening using an established FRED-Surflex consensus approach. *Chemical biology & drug design* 83(1):37-51.
105. Sperandio O, *et al.* (2014) Identification of novel small molecule inhibitors of activated protein C. *Thrombosis research*.
106. Chatzigeorgiou A, *et al.* (2014) Blocking CD40-TRAF6 signaling is a therapeutic target in obesity-associated insulin resistance. *Proceedings of the National Academy of Sciences of the United States of America* 111(7):2686-2691.
107. Vehar GA, *et al.* (1984) Structure of human factor VIII. *Nature* 312(5992):337-342.
108. Gitschier J, *et al.* (1984) Characterization of the human factor VIII gene. *Nature* 312(5992):326-330.
109. Lenting PJ, van Mourik JA, & Mertens K (1998) The life cycle of coagulation factor VIII in view of its structure and function. *Blood* 92(11):3983-3996.
110. Pemberton S, *et al.* (1997) A molecular model for the triplicated A domains of human factor VIII based on the crystal structure of human ceruloplasmin. *Blood* 89(7):2413-2421.
111. Saenko E, *et al.* (2001) Comparison of the properties of phospholipid surfaces formed on HPA and L1 biosensor chips for the binding of the coagulation factor VIII. *J Chromatogr A* 921(1):49-56.
112. Novakovic VA, *et al.* (2011) Membrane-binding properties of the Factor VIII C2 domain. *Biochem J* 435(1):187-196.
113. Liu Z, *et al.* (2010) Trp2313-His2315 of factor VIII C2 domain is involved in membrane binding: structure of a complex between the C2 domain and an inhibitor of membrane binding. *J Biol Chem* 285(12):8824-8829.
114. Lu J, Pipe SW, Miao H, Jacquemin M, & Gilbert GE (2011) A membrane-interactive surface on the factor VIII C1 domain cooperates with the C2 domain for cofactor function. *Blood* 117(11):3181-3189.
115. Takeshima K, Smith C, Tait J, & Fujikawa K (2003) The preparation and phospholipid binding property of the C2 domain of human factor VIII. *Thromb*

Haemost 89(5):788-794.
116. Farah CA & Sossin WS (2012) The role of C2 domains in PKC signaling. *Adv Exp Med Biol* 740:663-683.
117. Stace CL & Ktistakis NT (2006) Phosphatidic acid- and phosphatidylserine-binding proteins. *Biochim Biophys Acta* 1761(8):913-926.
118. McMullen BA, Fujikawa K, Davie EW, Hedner U, & Ezban M (1995) Locations of disulfide bonds and free cysteines in the heavy and light chains of recombinant human factor VIII (antihemophilic factor A). *Protein Sci* 4(4):740-746.
119. Wion KL, Kelly D, Summerfield JA, Tuddenham EG, & Lawn RM (1985) Distribution of factor VIII mRNA and antigen in human liver and other tissues. *Nature* 317(6039):726-729.
120. Hollestelle MJ, Geertzen HG, Straatsburg IH, van Gulik TM, & van Mourik JA (2004) Factor VIII expression in liver disease. *Thromb Haemost* 91(2):267-275.
121. Hollestelle MJ, *et al.* (2001) Tissue distribution of factor VIII gene expression in vivo--a closer look. *Thromb Haemost* 86(3):855-861.
122. Fay PJ, Coumans JV, & Walker FJ (1991) von Willebrand factor mediates protection of factor VIII from activated protein C-catalyzed inactivation. *J Biol Chem* 266(4):2172-2177.
123. Ngo JC, Huang M, Roth DA, Furie BC, & Furie B (2008) Crystal structure of human factor VIII: implications for the formation of the factor IXa-factor VIIIa complex. *Structure* 16(4):597-606.
124. Eaton D, Rodriguez H, & Vehar GA (1986) Proteolytic processing of human factor VIII. Correlation of specific cleavages by thrombin, factor Xa, and activated protein C with activation and inactivation of factor VIII coagulant activity. *Biochemistry* 25(2):505-512.
125. Bhopale GM & Nanda RK (2003) Blood coagulation factor VIII: An overview. *J Biosci* 28(6):783-789.
126. Lollar P & Parker ET (1991) Structural basis for the decreased procoagulant activity of human factor VIII compared to the porcine homolog. *J Biol Chem* 266(19):12481-12486.
127. Lamphear BJ & Fay PJ (1992) Proteolytic interactions of factor IXa with human factor VIII and factor VIIIa. *Blood* 80(12):3120-3126.
128. O'Brien DP, Johnson D, Byfield P, & Tuddenham EG (1992) Inactivation of factor VIII by factor IXa. *Biochemistry* 31(10):2805-2812.
129. Gilbert GE & Arena AA (1996) Activation of the factor VIIIa-factor IXa enzyme complex of blood coagulation by membranes containing phosphatidyl-L-serine. *J Biol Chem* 271(19):11120-11125.
130. Foster PA, Fulcher CA, Houghten RA, & Zimmerman TS (1990) A synthetic factor VIII peptide of eight amino acid residues (1677-1684) contains the binding region of an anti-factor VIII antibody which inhibits the binding of factor VIII to von Willebrand factor. *Thromb Haemost* 63(3):403-406.
131. Gilbert GE, Kaufman RJ, Arena AA, Miao H, & Pipe SW (2002) Four hydrophobic amino acids of the factor VIII C2 domain are constituents of both the membrane-binding and von Willebrand factor-binding motifs. *J Biol Chem* 277(8):6374-6381.
132. Saenko E, *et al.* (2001) Comparison of the properties of phospholipid surfaces formed on HPA and L1 biosensor chips for the binding of the coagulation factor VIII. *J Chromatogr A* 921(1):49-56.
133. Novakovic VA, *et al.* (2011) Membrane-binding properties of the Factor VIII C2

domain. *Biochem J* 435(1):187-196.
134. Liu Z, *et al.* (2010) Trp2313-His2315 of factor VIII C2 domain is involved in membrane binding: structure of a complex between the C2 domain and an inhibitor of membrane binding. *J Biol Chem* 285(12):8824-8829.
135. Zeerleder S, Hack CE, & Wuillemin WA (2005) Disseminated intravascular coagulation in sepsis. *Chest* 128(4):2864-2875.
136. Sarangi PP, Lee HW, & Kim M (2010) Activated protein C action in inflammation. *Br J Haematol* 148(6):817-833.
137. Popovic M, *et al.* (2012) Thrombin and vascular inflammation. *Molecular and cellular biochemistry* 359(1-2):301-313.
138. Sparkenbaugh EM, *et al.* (2014) Differential contribution of FXa and thrombin to vascular inflammation in a mouse model of sickle cell disease. *Blood* 123(11):1747-1756.
139. Flannery S & Bowie AG (2010) The interleukin-1 receptor-associated kinases: critical regulators of innate immune signalling. *Biochemical pharmacology* 80(12):1981-1991.
140. Wesche H, *et al.* (1999) IRAK-M is a novel member of the Pelle/interleukin-1 receptor-associated kinase (IRAK) family. *J Biol Chem* 274(27):19403-19410.
141. Croston GE, Cao Z, & Goeddel DV (1995) NF-kappa B activation by interleukin-1 (IL-1) requires an IL-1 receptor-associated protein kinase activity. *J Biol Chem* 270(28):16514-16517.
142. Cao Z, Henzel WJ, & Gao X (1996) IRAK: a kinase associated with the interleukin-1 receptor. *Science* 271(5252):1128-1131.
143. Muzio M, Ni J, Feng P, & Dixit VM (1997) IRAK (Pelle) family member IRAK-2 and MyD88 as proximal mediators of IL-1 signaling. *Science* 278(5343):1612-1615.
144. Suzuki N, Suzuki S, & Yeh WC (2002) IRAK-4 as the central TIR signaling mediator in innate immunity. *Trends in immunology* 23(10):503-506.
145. Janssens S & Beyaert R (2003) Functional diversity and regulation of different interleukin-1 receptor-associated kinase (IRAK) family members. *Mol Cell* 11(2):293-302.
146. Bolli R (2007) Preconditioning: a paradigm shift in the biology of myocardial ischemia. *American journal of physiology. Heart and circulatory physiology* 292(1):H19-27.
147. Huang J, Gao X, Li S, & Cao Z (1997) Recruitment of IRAK to the interleukin 1 receptor complex requires interleukin 1 receptor accessory protein. *Proceedings of the National Academy of Sciences of the United States of America* 94(24):12829-12832.
148. Wesche H, Henzel WJ, Shillinglaw W, Li S, & Cao Z (1997) MyD88: an adapter that recruits IRAK to the IL-1 receptor complex. *Immunity* 7(6):837-847.
149. Zhou H, *et al.* (2013) IRAK-M mediates Toll-like receptor/IL-1R-induced NFkappaB activation and cytokine production. *The EMBO journal* 32(4):583-596.
150. Su J, Zhang T, Tyson J, & Li L (2009) The interleukin-1 receptor-associated kinase M selectively inhibits the alternative, instead of the classical NFkappaB pathway. *Journal of innate immunity* 1(2):164-174.
151. Deng JC, *et al.* (2006) Sepsis-induced suppression of lung innate immunity is mediated by IRAK-M. *The Journal of clinical investigation* 116(9):2532-2542.
152. Hoogerwerf JJ, *et al.* (2012) Interleukin-1 receptor-associated kinase M-deficient mice demonstrate an improved host defense during Gram-negative pneumonia.

Molecular medicine 18:1067-1075.
153. van der Windt GJ, *et al.* (2012) Interleukin 1 receptor-associated kinase m impairs host defense during pneumococcal pneumonia. *The Journal of infectious diseases* 205(12):1849-1857.
154. Chassin C, *et al.* (2012) MicroRNA-146a-mediated downregulation of IRAK1 protects mouse and human small intestine against ischemia/reperfusion injury. *EMBO molecular medicine* 4(12):1308-1319.
155. Bamboat ZM, *et al.* (2010) Toll-like receptor 9 inhibition confers protection from liver ischemia-reperfusion injury. *Hepatology* 51(2):621-632.
156. Pope MR, Hoffman SM, Tomlinson S, & Fleming SD (2010) Complement regulates TLR4-mediated inflammatory responses during intestinal ischemia reperfusion. *Molecular immunology* 48(1-3):356-364.
157. Sun K, *et al.* (2012) Effect of taurine on IRAK4 and NF-kappa B in Kupffer cells from rat liver grafts after ischemia-reperfusion injury. *American journal of surgery* 204(3):389-395.
158. Hood L & Rowen L (2013) The human genome project: big science transforms biology and medicine. *Genome medicine* 5(9):79.
159. Lander ES, *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* 409(6822):860-921.
160. Fleischmann RD, *et al.* (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* 269(5223):496-512.
161. Zhu W, Lomsadze A, & Borodovsky M (2010) Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res* 38(12):e132.
162. Mills R, Rozanov M, Lomsadze A, Tatusova T, & Borodovsky M (2003) Improving gene annotation of complete viral genomes. *Nucleic Acids Res* 31(23):7041-7055.
163. Carvajal-Rodriguez A (2010) Simulation of genes and genomes forward in time. *Current genomics* 11(1):58-61.
164. Tatusova T, Ciufo S, Fedorov B, O'Neill K, & Tolstoy I (2014) RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic Acids Res* 42(1):D553-559.
165. Munch R, *et al.* (2005) Virtual Footprint and PRODORIC: an integrative framework for regulon prediction in prokaryotes. *Bioinformatics* 21(22):4187-4189.
166. Carlson JM, Chakravarty A, DeZiel CE, & Gross RH (2007) SCOPE: a web server for practical de novo motif discovery. *Nucleic Acids Res* 35(Web Server issue):W259-264.
167. Zimmer JS, Monroe ME, Qian WJ, & Smith RD (2006) Advances in proteomics data analysis and display using an accurate mass and time tag approach. *Mass spectrometry reviews* 25(3):450-482.
168. Lepoittevin C, *et al.* (2010) In vitro vs in silico detected SNPs for the development of a genotyping array: what can we learn from a non-model species? *PloS one* 5(6):e11034.
169. Bromberg Y & Rost B (2008) Comprehensive in silico mutagenesis highlights functionally important residues in proteins. *Bioinformatics* 24(16):i207-212.
170. Arvestad L, Berglund AC, Lagergren J, & Sennblad B (2003) Bayesian gene/species tree reconciliation and orthology analysis using MCMC. *Bioinformatics* 19 Suppl 1:i7-15.
171. McGuffee SR & Elcock AH (2010) Diffusion, crowding & protein stability in a

dynamic molecular model of the bacterial cytoplasm. *PLoS computational biology* 6(3):e1000694.
172. Frembgen-Kesner T & Elcock AH (2010) Absolute protein-protein association rate constants from flexible, coarse-grained Brownian dynamics simulations: the role of intermolecular hydrodynamic interactions in barnase-barstar association. *Biophysical journal* 99(9):L75-77.
173. Wang Y, Hammes F, De Roy K, Verstraete W, & Boon N (2010) Past, present and future applications of flow cytometry in aquatic microbiology. *Trends in biotechnology* 28(8):416-424.
174. Stahura FL & Bajorath J (2005) New methodologies for ligand-based virtual screening. *Current pharmaceutical design* 11(9):1189-1202.
175. Villoutreix BO, *et al.* (2007) Free resources to assist structure-based virtual ligand screening experiments. *Current protein & peptide science* 8(4):381-411.
176. Gopalakrishnan K, Sowmiya G, Sheik SS, & Sekar K (2007) Ramachandran plot on the web (2.0). *Protein and peptide letters* 14(7):669-671.
177. Colovos C & Yeates TO (1993) Verification of protein structures: patterns of nonbonded atomic interactions. *Protein Sci* 2(9):1511-1519.
178. Bowie JU, Luthy R, & Eisenberg D (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science* 253(5016):164-170.
179. Luthy R, Bowie JU, & Eisenberg D (1992) Assessment of protein models with three-dimensional profiles. *Nature* 356(6364):83-85.
180. Landgraf R, Xenarios I, & Eisenberg D (2001) Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *Journal of molecular biology* 307(5):1487-1502.
181. Pontius J, Richelle J, & Wodak SJ (1996) Deviations from standard atomic volumes as a quality measure for protein crystal structures. *Journal of molecular biology* 264(1):121-136.
182. Berjanskii M, *et al.* (2010) PROSESS: a protein structure evaluation suite and server. *Nucleic Acids Res* 38(Web Server issue):W633-640.
183. Hooft RW, Vriend G, Sander C, & Abola EE (1996) Errors in protein structures. *Nature* 381(6580):272.
184. Vaguine AA, Richelle J, & Wodak SJ (1999) SFCHECK: a unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model. *Acta crystallographica. Section D, Biological crystallography* 55(Pt 1):191-205.
185. Kleywegt GJ, *et al.* (2004) The Uppsala Electron-Density Server. *Acta crystallographica. Section D, Biological crystallography* 60(Pt 12 Pt 1):2240-2249.
186. Wallner B & Elofsson A (2003) Can correct protein models be identified? *Protein Sci* 12(5):1073-1086.
187. Melo F & Feytmans E (1998) Assessing protein structures with a non-local atomic interaction energy. *Journal of molecular biology* 277(5):1141-1152.
188. Willard L, *et al.* (2003) VADAR: a web server for quantitative evaluation of protein structure quality. *Nucleic Acids Res* 31(13):3316-3319.
189. Laskowski RA (2009) PDBsum new things. *Nucleic Acids Res* 37(Database issue):D355-359.
190. Neves MA, Totrov M, & Abagyan R (2012) Docking and scoring with ICM: the benchmarking results and strategies for improvement. *Journal of computer-aided molecular design* 26(6):675-686.
191. Vilar S, Cozza G, & Moro S (2008) Medicinal chemistry and the molecular

operating environment (MOE): application of QSAR and molecular docking to drug discovery. *Current topics in medicinal chemistry* 8(18):1555-1572.
192. Krieger E, Koraimann G, & Vriend G (2002) Increasing the precision of comparative models with YASARA NOVA--a self-parameterizing force field. *Proteins* 47(3):393-402.
193. Case DA, *et al.* (2005) The Amber biomolecular simulation programs. *Journal of computational chemistry* 26(16):1668-1688.
194. Brooks BR, *et al.* (2009) CHARMM: the biomolecular simulation program. *Journal of computational chemistry* 30(10):1545-1614.
195. Pronk S, *et al.* (2013) GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* 29(7):845-854.
196. Nielsen M, Lundegaard C, Lund O, & Petersen TN (2010) CPHmodels-3.0--remote homology modeling using structure-guided sequence profiles. *Nucleic Acids Res* 38(Web Server issue):W576-581.
197. Zhou H & Skolnick J (2007) Ab initio protein structure prediction using chunk-TASSER. *Biophysical journal* 93(5):1510-1518.
198. Bradley P, Misura KM, & Baker D (2005) Toward high-resolution de novo structure prediction for small proteins. *Science* 309(5742):1868-1871.
199. Simons KT, Kooperberg C, Huang E, & Baker D (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *Journal of molecular biology* 268(1):209-225.
200. Pieper U, *et al.* (2014) ModBase, a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res* 42(1):D336-346.
201. Remmert M, Biegert A, Hauser A, & Soding J (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature methods* 9(2):173-175.
202. Kelley LA & Sternberg MJ (2009) Protein structure prediction on the Web: a case study using the Phyre server. *Nature protocols* 4(3):363-371.
203. Arnold K, Bordoli L, Kopp J, & Schwede T (2006) The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* 22(2):195-201.
204. Sali A & Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. *Journal of molecular biology* 234(3):779-815.
205. Lambert C, Leonard N, De Bolle X, & Depiereux E (2002) ESyPred3D: Prediction of proteins 3D structures. *Bioinformatics* 18(9):1250-1256.
206. Krivov GG, Shapovalov MV, & Dunbrack RL, Jr. (2009) Improved prediction of protein side-chain conformations with SCWRL4. *Proteins* 77(4):778-795.
207. L. DW (2007) The PyMOL molecular graphics system. *USA: DeLano Scientific*.
208. Humphrey W, Dalke A, & Schulten K (1996) VMD: visual molecular dynamics. *Journal of molecular graphics* 14(1):33-38, 27-38.
209. Zhou HX & Shan Y (2001) Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins* 44(3):336-343.
210. Singh R, Park D, Xu J, Hosur R, & Berger B (2010) Struct2Net: a web service to predict protein-protein interactions using a structure-based approach. *Nucleic Acids Res* 38(Web Server issue):W508-515.
211. Bradford JR & Westhead DR (2005) Improved prediction of protein-protein binding sites using a support vector machines approach. *Bioinformatics* 21(8):1487-1494.

212. McGann M (2011) FRED pose prediction and virtual screening accuracy. *Journal of chemical information and modeling* 51(3):578-596.
213. Jain AN (2003) Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *Journal of medicinal chemistry* 46(4):499-511.
214. Liebeschuetz JW, Cole JC, & Korb O (2012) Pose prediction and virtual screening performance of GOLD scoring functions in a standardized test. *Journal of computer-aided molecular design* 26(6):737-748.
215. Ewing TJ, Makino S, Skillman AG, & Kuntz ID (2001) DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *Journal of computer-aided molecular design* 15(5):411-428.
216. Friesner RA, *et al.* (2006) Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *Journal of medicinal chemistry* 49(21):6177-6196.
217. Zsoldos Z, Reid D, Simon A, Sadjad BS, & Johnson AP (2006) eHiTS: an innovative approach to the docking and scoring function problems. *Current protein & peptide science* 7(5):421-435.
218. McGaughey GB, *et al.* (2007) Comparison of topological, shape, and docking methods in virtual screening. *Journal of chemical information and modeling* 47(4):1504-1519.
219. Venkatachalam CM, Jiang X, Oldfield T, & Waldman M (2003) LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. *Journal of molecular graphics & modelling* 21(4):289-307.
220. Koes DR & Camacho CJ (2011) Pharmer: efficient and exact pharmacophore search. *Journal of chemical information and modeling* 51(6):1307-1314.
221. Corbeil CR, Englebienne P, & Moitessier N (2007) Docking ligands into flexible and solvated macromolecules. 1. Development and validation of FITTED 1.0. *Journal of chemical information and modeling* 47(2):435-449.
222. Kramer B, Rarey M, & Lengauer T (1999) Evaluation of the FLEXX incremental construction algorithm for protein-ligand docking. *Proteins* 37(2):228-241.
223. Trott O & Olson AJ (2010) AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry* 31(2):455-461.
224. An J, Totrov M, & Abagyan R (2005) Pocketome via comprehensive identification and classification of ligand binding envelopes. *Molecular & cellular proteomics : MCP* 4(6):752-761.
225. Laurie AT & Jackson RM (2005) Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics* 21(9):1908-1916.
226. Huang B (2009) MetaPocket: a meta approach to improve protein ligand binding site prediction. *Omics : a journal of integrative biology* 13(4):325-330.
227. Volkamer A, Kuhn D, Rippmann F, & Rarey M (2012) DoGSiteScorer: a web server for automatic binding site prediction, analysis and druggability assessment. *Bioinformatics* 28(15):2074-2075.

Chapter 2

Optimization of compound ranking for structure-based virtual ligand screening using an established FRED-Surflex consensus approach

Running Title: Optimized scoring for hierarchic ensemble docking

Jiangfeng Du, Ivo W. M. Bleylevens, Albert V. Bitorina,
Kanin Wichapong, Gerry A.F. Nicolaes

Chem Biol Drug Des. 83(1):37-51, Jan, 2014

Abstract

Use of multiple target conformers has been applied successfully in virtual screening campaigns; however a study on how to best combine scores for multiple targets in a hierarchic method that combines rigid and flexible docking is not available. In this study, we used a dataset of 59,479 compounds to screen multiple conformers of four distinct protein targets to obtain an adapted and optimized combination of an established hierarchic method that employs the programs FRED and Surflex. Our study was extended and verified by application of our protocol to ten different data sets from the directory of useful decoys (DUD). We quantitated overall method performance in ensemble docking and compared several consensus scoring methods to improve the enrichment during virtual ligand screening. We conclude that one of the methods used, which employs a consensus weighted scoring of multiple target conformers, performs consistently better than methods that do not include such consensus scoring. For optimal overall performance in ensemble docking, it is advisable to first calculate a consensus of FRED results and use this consensus as a sub-dataset for Surflex screening. Furthermore we identified an optimal method for each of the chosen targets and propose how to optimize the enrichment for any target.

Keywords virtual ligand screening, SB-VLS, ensemble docking, consensus scoring, drug design

Introduction

Together with chemical synthesis and combinatorial methods, rational design and high-throughput methods have become essential to modern drug discovery (1, 2). Moreover, the currently more than 20 million commercially available compounds, representing a reasonable diversity of the chemical space, greatly enhance the use of *in silico* screening methods (3). In fact, the first drugs that have been discovered through *in silico* techniques are now reaching the markets (4, 5). In virtual screening, small molecules are identified that are complementary in shape and charge to a biomolecular target to which they are intended to interact. Two kinds of virtual ligand screening methods in rational drug design can be distinguished: ligand-based methods and structure-based methods. In ligand-based methods, molecular descriptors, 1D, 2D or 3D structural information of active compounds are used to select other compounds from an *in silico* collection of compounds (5, 6). This selection process is performed by various methods that include: similarity and substructure searching (7), pharmacophore matching (8), 3D shape matching (9), electrostatic distribution similarity search (10) and shape-based screening approach such as ROCS (11) , Ultrafast Shape Recognition (USR) (12) and Phase Shape (13). Besides having an accurate search algorithm, the ligands need to be in comprehensive conformations and have correct protonation states to perform a successful ligand based screening (14).

Structure-based virtual ligand screening on the other hand is used when the structure of the target protein is known or can be modeled

(15). This enables the molecular docking of small molecules into a defined binding pocket of the target. Generation and consecutive scoring of the binding pose for each compound from a large 3D database of compounds yields a ranked list of potential ligands bound to the target. Structure-based virtual ligand screening (SBVLS) methods can be generally divided into two different docking methods: rigid and flexible. During rigid body docking, both ligand and receptor are treated as rigid objects that cannot change their spatial shape during the docking process. Rigid body-based methods require relatively little intensive computations (16, 17). Using rigid body docking, numerous successful experiments have been performed e.g. refs (18, 19). During flexible docking, the conformations of the ligand and/or the receptor are altered or generated during the docking process to fit its docking partner. In recent years, flexible docking has been applied successfully in many cases (20, 21), despite the large number of degrees of freedom that has to be considered.

Many software packages, both commercial and non-commercial, are used for virtual ligand screening (VLS); some examples are FRED (22, 23), Surflex (24) , Glide (25, 26), GOLD (27), ICM (28), and LigandFit (29), but many more are available (30). Each of these packages have their advantages and disadvantages, more details of each molecular docking program can be found in a recent review by Yuriev (31). Even though many docking new programs are being developed, most docking programs still have common problems, for example with protein flexibility, the solvation effect, and the treatment of waters or metal at the binding pocket. Moreover, binding modes

and binding affinities of compounds are predicted based on docking scores which mostly are a simple linear equation generated by using focused data sets. Although force field-based scoring function is applied in some docking programs, the entropy change upon ligand binding is not included in the scoring function. Furthermore it is often found that there is no correlation between docking score and binding affinities of compound. Therefore, some limitations for the application of molecular docking in the drug discovery process still exist and a general method that performs optimal on every target is however presently not at hand. Inevitably concessions are made in each approach to find a balance between the accuracy of the method and the required CPU time consumption when screening large collections of virtual molecules.

To approach the docking problem pragmatically, multi-step VLS protocols have been applied in which consecutive screenings are used to narrow down a large database of molecules. One such fast multi-step protocol by Miteva et al.(32) has been applied with considerable success (33, 34). The protocol allows fast screening by a combination of rigid body docking using FRED with two flexible docking tools, DOCK and Surflex, to obtain overall better results than with application of any of the docking tools individually. However, how to best combine both methods is a question that has not sufficiently been addressed yet and which is systematically dealt with in this paper.

Much like their ligands, proteins are dynamic by nature and can adopt different conformations (35, 36). Therefore, in a SBVLS

approach, a single conformation of a receptor might not be representative for the encounter conformation of a receptor that binds a given ligand. Thus, multiple conformations of a target can be selected for a SBVLS study (37, 38) at the expense of the introduction of a larger total number of degrees of freedom and a significant increase of the computational burden (39, 40). In ensemble docking, different conformations of the target protein can be obtained from different X-ray structures (41)-(42), from snapshots of a molecular dynamics simulation (43) or from normal mode analysis (44). To integrate results for different conformations, a consensus scoring must be applied. A good consensus score function is able to reduce the number of false positive docking results and increase the predicted binding affinity in a structure-based virtual ligand screening (45, 46). Previous works have explored the optimal ensemble size, 4-5, from crystallographic structures and demonstrated the optimal structures used for ensemble should be those with the largest ligands in their pockets (36, 47, 48). In this work however, we choose to use structures from normal mode analysis, which have somewhat different features than X-ray structures (49). Therefore we have investigated the ensemble size from normal modes (see Figure 1 and Figure 2) and chose to generate eight conformers for each target to screen the small molecule dataset in order to sample binding pocket flexibility. Application of such ensemble docking however implies that one has to decide on how to optimally combine the results from multiple conformers into the multi-step VLS protocol.

We have performed an *in silico* study, using a dataset of 59,479

compounds that was seeded with 289 actives, retrieved from the directory of useful decoys (DUD) (50) (directory of useful decoys, http://dud.docking.org/) and extended and verified the study by application of our protocol to ten different data sets from the DUD.

With this study we have addressed the following research questions:

1. Is the use of multiple protein conformations better than the use of a single protein conformation for the combined FRED-Surflex virtual ligand screening procedure used here?

2. How to process those multiple protein conformations to obtain an overall better enrichment for structurally diverse protein targets, while applying a consensus docking score to optimally combine a rigid and a flexible docking programs (FRED and Surflex) respectively?

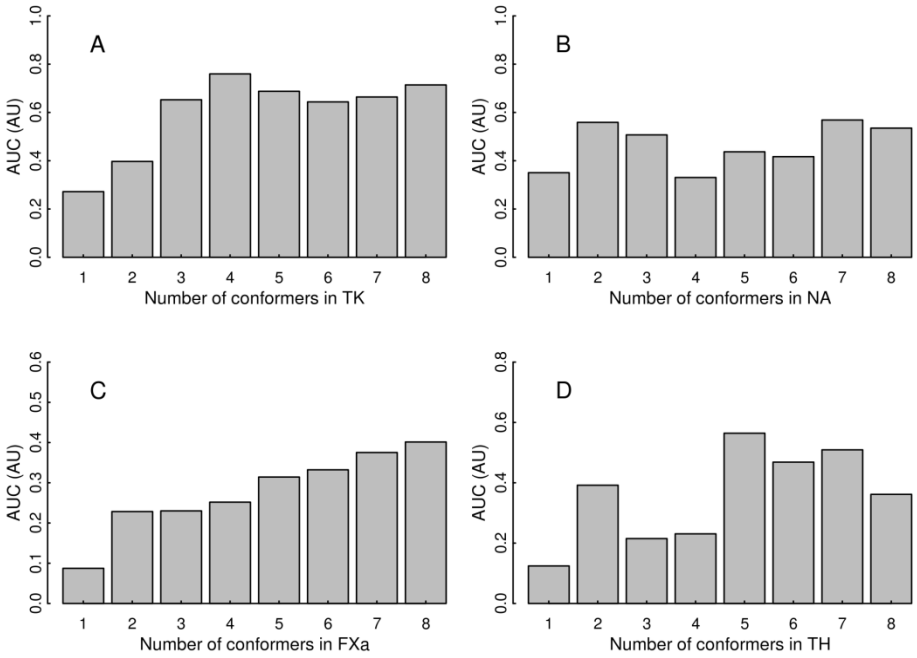


Figure 1. Analysis of the combined FRED-Surflex VLS by including different numbers of conformers for each target and with the decoys retrieved from the Chembridge database. The performances are indicated by the area under the ROC curves at 5% FPR fraction when different numbers (*i*) of target conformations are used. X-axis indicates the numbers (*i*) of conformers used in VLS (0<*i*<9). Y-axis, the weighted AUC value indicates the virtual screening performance at 5% FPR fraction for each ensemble size. Four targets have been tested in this experiment (A: *TK*, B: *NA*, C: *FXa*, D: *TH*). Consensus method *concom* was used with the S_{mean} method.

Methods

Target Selection and Preparation

Four different protein targets were selected according to the diversity of their binding-site properties (Table 1), the availability of a high resolution protein-small molecule complex structure and the availability of a number of confirmed ligands to the intended target binding pocket. The 3D crystal structures of thymidine kinase (*TK*, PDBID: 3F0T), neuraminidase (*NA,* 3O9K), coagulation factor Xa (*FXa*, 2XBV), and thrombin (*TH*, 2ZFQ) were retrieved from the PDB-redo database (http://www.cmbi.ru.nl/pdb_redo/) and ligands were removed. To generalize the potential binding mode and mimic a situation of compound selection with no prior knowledge of the mode of binding or of a role of waters in binding, we removed any crystal water molecules. This step is supported by the fact that waters in the binding pockets are not conserved in the crystallographic complexes for the proteins considered for this study (e.g. thrombin structures 2ZFQ.pdb and 3U9A.pdb).

Proteins were prepared by using the ICM-convert module implemented in with ICM-Pro (Version 3.4.6) (51). Hydrogen atoms were added to the protein structures according to the protonation

state at pH=7 which means that all acidic residue (Asp and Glu) are deprotonated and all basic residue (Lys and Arg) are protonated. Next, eight different conformers were generated for each target using the software NORMA (49) to apply a simplex simulated annealing minimization of the four targets. Subsequently, normal mode analysis (NMA) (52) was used to generate a list of 'large' conformationally changed modes for *TK, NA, FXa* and *TH* which were selected based on the degree of collectivity of movement, and the overlaps of conformations (53) and the range of RMSD values between different conformations of each protein are summarized in Table 2. The number of eight conformers was chosen after an initial analysis in which we determined for a varying number of conformers for each of the targets, at the initial percentage of the compound database (0-5%), the corresponding enrichments and whether inclusion of additional conformers would result in overall improvement or not (Figures 1 and 2). The pockets to which ligands were bound in the co-crystallized complexes were selected as target pockets for virtual ligand screening using FRED and Surflex and the pocket volume, docking inner and outer contours and mutable residues were calculated (see Table 1).

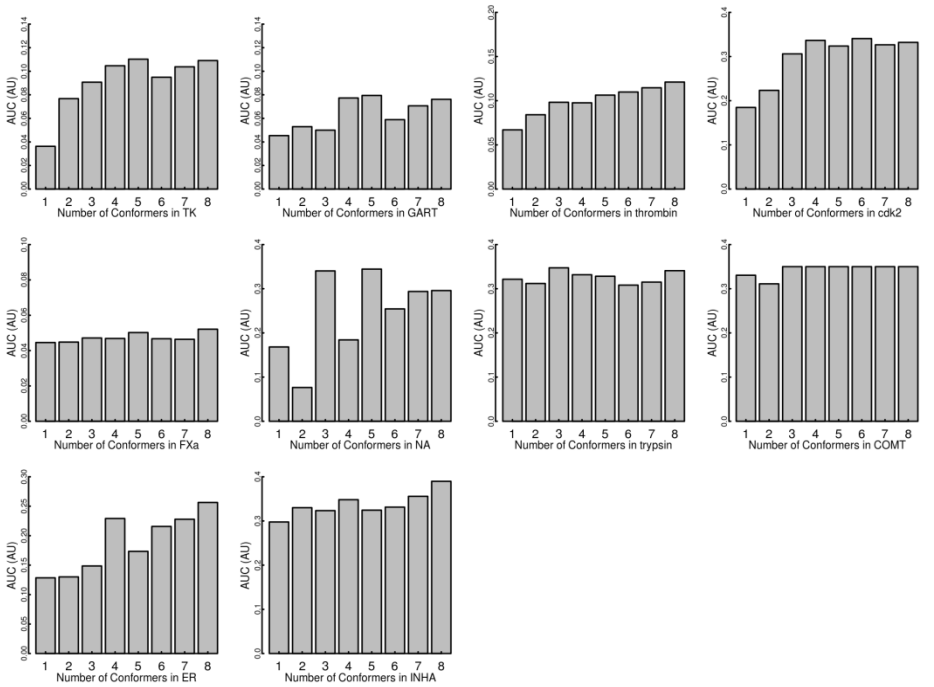


Figure 2. Analysis of the combined FRED-Surflex VLS by including different numbers of conformers for each target and with DUD decoys. The performances are indicated by the area under the ROC curves at 5% FPR fraction of each VLS when different numbers (*i*) used. X-axis indicates the numbers (*i*) of conformers used in VLS (0<*i*<9). On the Y-axis, the weighted AUC value indicates the virtual screening performance at 5% FPR fraction for each ensemble size. Ten targets have been tested in this experiment. Consensus method *concom* was used with the S_{mean} method.

Compound Dataset

Active ligands of *TK, NA, FXa* and *TH* were retrieved from the Directory of Useful Decoys (DUD, Version 2) (54). The number of active ligands used for each of the targets was 22, 49, 146, and 72 respectively, and these ligands can be clustered in several groups based on their chemical diversity (Table 3). Since the numbers of decoys available from DUD for some targets is relatively small (e.g 891 decoys in *TK*), and we wished to mimic as well as possible a real

virtual screening campaign we stochastically retrieved compounds from the ChemBridge database, December 2010 (55). Next, we verified the quality of the decoy data set and we found no overlap of compounds between the active and decoy data sets. The decoy data set contains a total amount of 59,479 (~10% of ChemBridge database) unique molecules. We verified for several key physico-chemical molecular properties that the dataset and known actives are similar (see Figure 3). Since one of the aims of our work is to compare the performance of virtual screening when using single and multiple conformations of protein, we used the same database for all screening steps.

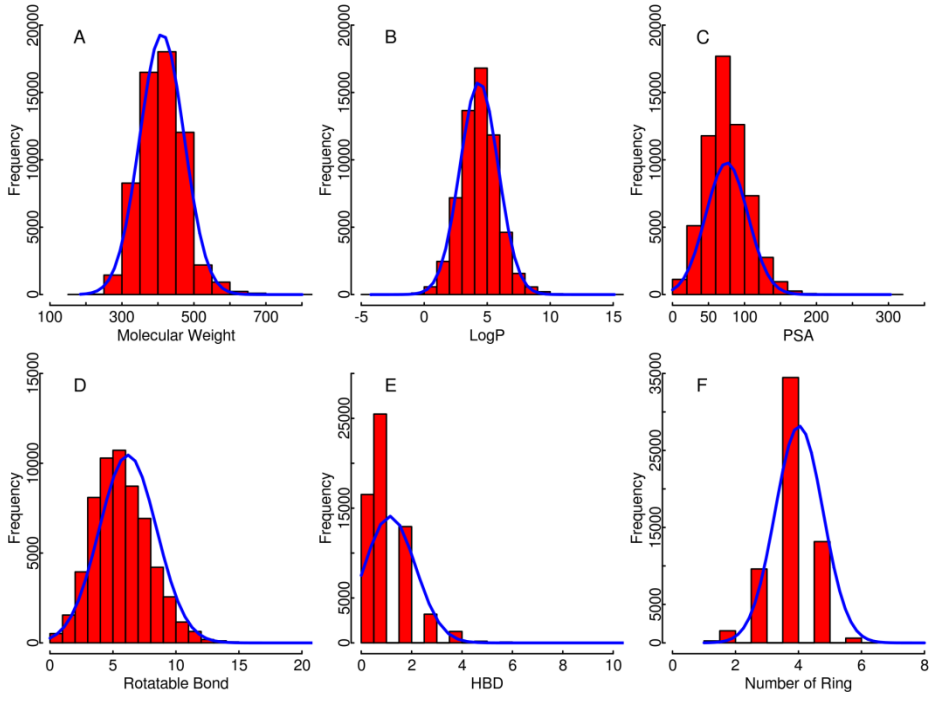


Figure 3. Physico-chemical properties of the Chembridge dataset selected in this work. 59,479 compounds were analyzed by their molecular weight (a), partition coefficient (logP, b), polar surface area (PSA, c), rotatable bonds (d),

hydrogen bond donors (HBD, e) and Rings (f). The curves in each sub-figure are the probability density function of the active ligands and the bars indicate the numbers of decoy compounds in the given range (x-axis). The figure was generated by R.

The program Omega (ver2.1.0) (16) was used to convert all compounds into 3D multi-conformers and to add hydrogen atoms/Gasteiger partial charges. For rigid body docking with FRED, the generation of multi-conformer structures for each compound was required to sample the conformational space for the small molecules. Omega was run with the *RMSD* threshold value set to 0.8 Å, the energy window set to 10.0 kcal/mol and a maximum conformations per compound of 50. These settings resulted in a total of 1,128,676 structures with an average of 19 conformations per unique compound.

We used 4 different data sets (active compounds *TK, NA, FXa* and *TH* from the DUD combined with decoys selected from ChemBridge database) to mimic a real virtual screening campaign. However, to substantiate our finding that multiple conformations and a combined FRED-Surflex docking approach can result in better overall performance as compared to use of a single conformation and using FRED or Surflex alone, we extended our study by inclusion of actives and decoys from the well-defined and unbiased DUD data set for *TK, NA, FXa* and *TH* as well as 6 additional target proteins (*GART* - glycinamide ribonucleotide transformylase, *CDK2* - Cyclin dependent kinase 2, *Trypsin*, *COMT* - Catechol O-methyltransferase, *ER* - Estrogen receptor and *INHA* - Enoyl ACP reductase).

Rigid body docking

FRED (Fast Rigid Exhaustive Docking Open Eye Scientific Software Inc., Version 2.2.5) (22, 23, 56, 57) is a rigid protein-ligand docking program which can generate an ensemble pose, and is able to rank and optimize these ensembles by means of a scoring function. During the docking process, FRED treats all molecules as rigid body objects which cannot change their spatial shape. In this work, binding-sites were selected on basis of the crystal structures from the PDB-redo database (58) (http://www.cmbi.ru.nl/pdb_redo/). The binding-site boxes were detected by using the "FRED_receptor" module with the molecule site detection method and a site shape potential was generated with medium quality control (22). Detailed binding-site information of the four targets' conformers are given in Table 1. The multi-conformer compound dataset was docked onto each of the eight conformers per target and the docked poses were scored using FRED's scoring functions, *OEChemscore* (59) and *Shapegauss* (22). FRED further generated a default consensus score of *OEChemscore* and *Shapegauss*. Each score was generated with the 3D conformation of the corresponding ligand-receptor complex as well as of the undocked compounds.

Flexible docking

Surflex (BioPharmics LLC, Version 2.514) relies on a surface-based molecular similarity using a re-parameterized empirical scoring function (24). Since however the overall receptor conformation is essentially unchanged in Surflex docking, we used multiple

conformations of the four target proteins (*TK, NA, FXa,* and *TH*) during Surflex docking. In this study, we selected the residues surrounding a binding-site (see Table 1) to define the protomol, the pseudo-molecule which serves as the target in Surflex. The optimal protomol should well describe the pocket shape and can be varied by the number of atomic probes (Proto_bloat) and the degree of buriedness with the surrounding residues (Proto_thresh); the parameters *proto_thresh and proto_bloat* together determine the quality of a protomol. Taking into account the diverse properties of the four target pockets, *proto_thresh* was set to 0 and *proto_bloat* to 0 for the buried pocket of *TK*. For the deeper pockets of *FXa* and *NA, proto_thresh* was set to 0.5 and proto_bloat to 1. Since thrombin has a relatively open/shallow pocket, *proto_thresh* was chosen as 1 and proto_bloat as 3. Protomols were visualized with the PyMOL Molecular Graphics System (Version 1.2R1) (60) to ensure proper coverage of the desired binding-site area. Parameters of *value1* and *value2* of each target in the command line for the protomol generation: "*Surflex proto_thresh value1 proto_bloat value2 resproto residue-list protein.mol2 log* were obtained through prior optimization experiments. The complete set of compounds (59,479 compounds) was docked using this method.

Docked poses were generated and scored as described (61) and pose scores were post-processed using thresholds of 1.0 for polarity, of 100 for rotation and a crash threshold of -1.0. This resulted in 8 docking lists per target.

Data Analysis

Prerequisites Definitions

Using FRED and Surflex, two times eight ranked docking lists were produced for each target, which were further processed in Matlab (Version R2010A) and R (62). Out of the multiple conformations of a single compound in each docking list generated by FRED, only the conformer with the best score was selected for further processing and the rest were discarded. The threshold for inclusion for further processing was treated as an unknown variable, see also below. We addressed the variability of compound scores for the different target conformations and the possibility of outlier scores to influence a consensus score for a compound. Seven consensus methods, called S_{Mean}, $S_{TrimMean}$, S_1, S_2, S_3, S_4, and S_{opt}, were used in this study to identify an optimal consensus scoring protocol. The different consensus scoring methods were defined as follows,

$S_{Mean}(Q) = \mu = (v_1+v_2+v_3+v_4+v_5+v_6+v_7+v_8)/8$

$S_{TrimMean}(Q) = \hat{u} = (v_1+v_2+v_3+v_4+v_5+v_6)/6$, where v_7, v_8 are the highest and lowest in V_i

$S_1(Q) = \mu/\sigma$

$S_2(Q) = \mu/\sigma + \mu$

$S_3(Q) = \mu - \sigma$

$S_4(Q) = \mu + \sigma$

$S_{opt}(Q) = p_1*v_1+p_2*v_2+p_3*v_3+p_4*v_4+p_5*v_5+p_6*v_6+p_7*v_7+p_8* v_8+ p_9*\sigma$

Where:
v_i is the ranked value of a compound which is docked in a target conformer i,
Q is any compound of the dataset,
σ is the standard deviation of $v_1, \ldots, v_i$,
μ is the mean/average of the scores $v_1, \ldots, v_i$ of one compound,
$\hat{u}$ is the 25% trimmed mean of the scores $v_1, \ldots, v_i$ of one compound (i.e., leaving out the highest and lowest score when computing the mean in the case of 8 scores),

S is the final score of one compound after weighting the compound ranked values in different protein conformers.

The parameters p_1-p_9 in the formula of S_{opt} above, are computed from the results of a mathematical optimization problem. Here, the goal is to find values p_1 to p_9 that maximize the performance defined in terms of the enrichments of both FRED and Surflex simultaneously. The Levenberg-Marquardt algorithm of the Fminsearch function in Matlab is applied to generate the parameters for S_{opt}. This algorithm implements an iterative unconstrained technique that locates the minimum of a function that is expressed as the sum of squares of nonlinear functions. One could see the S_{opt} consensus list as a weighted mean consensus list, where the standard deviation is also taken into account (for the normal mean S_{Mean} it holds that p_1-p_8=1, p_9=0). Note that one should recompute the values p_1-p_9 for every experiment since they depend on the actual values v_i, a procedure that takes a few seconds.

Comparison of FRED and Surflex docking performance

The enrichment factor (EF), defined as the number of active compounds found in a certain percentage of a dataset can be calculated to assess the quality of the virtual screening method (63). The EF was computed at the 1% of the ranked database, and the highest EF at this considered point is equal to 100 indicating that all active compounds are obtained.

Moreover, the performance of each virtual screening and scoring approach was investigated by plotting the receiver operator characteristic (ROC) curves which show the ability to discriminate

known-active compounds from decoys. Docking results were sorted according to the docking score of the compounds and consequently the true positive rate *(TPR)* and the false positive rate *(FPR)* were computed using the equations below;

$$TPR = \frac{TP}{N < \text{actives} >} \quad \text{and} \quad FPR = \frac{FP}{N <\text{decoys}>} ,$$

where *N<actives>* and *N<decoys>* represent the number of known actives and decoys. Docking solutions that have a better or equal score to a particular compound are considered as positive solutions. Thus, active compounds at the considered point are true positive (*TP*) whereas decoys at the same point are false positive (*FP*). The perfect ROC curve should show a steep rise at the beginning to reach the maximum y value and then continue parallel to the x-axis which means that all the known-active compounds are recovered during early screening. Moreover, the area under the ROC curve (AUC) of each docking and scoring method was calculated. AUC represents the overall quality of each screening/scoring method and also indicates the probability to rank a randomly chosen active higher than a randomly chosen inactive.

A statistic p-values matrix was applied in the study to evaluate the differences between 7 consensus scoring methods for all the protein targets in FRED, Surflex and combined FRED-Surflex. The calculation of p-values (41) can be described as follows:

$$p = erfc \frac{|\Delta AUC|}{\sqrt{(2)} * SE_{\Delta}}$$

here |ΔAUC| is the absolute of the AUC values from any two scoring methods. Where SEΔ is derived from formula

$$SE_{\Delta} = \sqrt{\frac{Var_{\Delta a}}{N_{actives}} + \frac{Var_{\Delta a}}{N_{decoys}}}$$

where $N_{actives}$ and N_{decoys} are the number of active ligands and decoys respectively in a dataset, while $Var_{\Delta a}$, and $Var_{\Delta d}$ are derived from formula,

$$Var_{\Delta a} = \frac{1}{N_{actives}} * \sum \left(\left(FPRi,a - FPRi,b \right) - \Delta AUC \right)^2$$

$$Var_{\Delta d} = \frac{1}{N_{decoys}} * \sum \left(\left(TPRi,a - TPRi,b \right) - \Delta AUC \right)^2$$

Where TPR_i is the true positive rate at decoy position i and FPR_i is the false positive rate at the active position i. With the statistical p-value, the virtual screening performance of any two methods can be quantified to be similar or different. Two methods are treated as statistically different when the p-value <0.05 in a 95% confidence interval (CI), or the p-value < 0.1 in a 90% CI. For both FRED and Surflex, we calculated the enrichment lists for every conformer of the four protein targets, resulting in four times eight lists from which consensus lists (by means of methods S_{Mean}, $S_{TrimMean}$, S_1, S_2, S_3, S_4, and S_{opt}) for the four protein targets were calculated. In order to yield the combined FRED-Surflex enrichments, we used two different methods (see also Figure 4) to combine the results from FRED and Surflex. As described, eight ranked compound lists are generated by FRED docking scores for each protein conformer. The FRED ranked lists LI-L8 of conformer i are to be used as the input dataset for subsequent Surflex docking of the corresponding conformer i, which

will produce eight combined FRED-Surflex ranked lists called SDL_i ($0<i<9$). This method will be called the separation combination (*sepcom*).

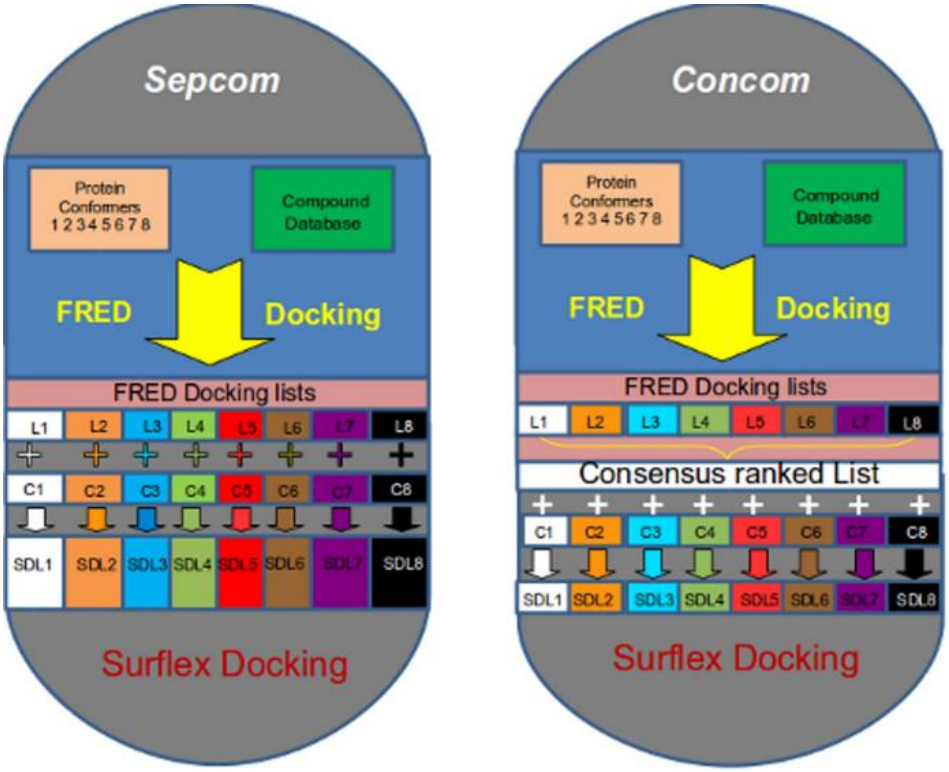


Figure 4. A flow chart depicting two different methods to the combined FRED-Surflex VLS. Left (*Sepcom*): FRED screened lists (L1, L2, ..., L8) for each conformer (C1, C2, ..., C8) are used as input compounds dataset only for its relevant conformer in subsequence Surflex screening, and then eight Surflex screened lists (SDL1, ..., SDL8) will be produced. **Right (*concom*):** consensus list is calculated from FRED docking lists (L1, L2, ..., L8), and it is then used as the input compounds dataset for all the eight conformers in the subsequence Surflex.

Alternatively, the eight FRED ranked lists LI-L8 are first combined into a consensus ranked list (via one of the consensus methods described above) and then the consensus list serves as an input dataset for each of the eight conformers in subsequent Surflex docking; resulting in eight combined FRED-Surflex ranked lists called

SDL_i. This latter method we called the consensus combination (*concom*).

Stepwise improvement of FRED and Surflex combination

The amount of compounds commonly selected after FRED docking that is used for consecutive Surflex docking (typically 30-60 % of the initial scores) in our VLS protocol is based on empirical observation (32). In order to further rationalize this percentage, we measured the stepwise improvement of the FRED and Surflex combination by variation of the amount of compounds that is allowed to proceed from FRED to Surflex. If the top X1 % ranked FRED result and top Y1 % ranked Surflex list contain a number of 'A' active ligands, and in the top X2 % (FRED) and top Y2 % (Surflex) of the total ranked list contains a number of 'A+1' actives, then, the stepwise improvement Z is defined as:

$$Z=((X2-X1)- (Y2-Y1)) / (X2- X1) *100\% \quad (\text{with } 0< X1<X2<100 \text{ and } 0<Y1<Y2 <100)$$

Z is defined as a value to describe the improvement of each step by Surflex. A positive Z score indicates that Surflex improves overall enrichment at this particular percentage. Improvement curves (Z curves) were generated by application of the seven consensus methods mentioned above (S_{Mean}, $S_{TrimMean}$, S_1, S_2, S_3, S_4, and S_{opt}) for the combined FRED-Surflex in the four targets.

Results

Pocket Characteristics

Pocket characteristics, average CPU times for the four different targets are shown in Table 2. A correlation between the pocket volume and calculation time was observed for TK, FXa and TH when FRED was used (Pearson correlation coefficient of 0.52; 0.18; 0.97 and 0.79 respectively), no such correlation was observed for any of the targets when Surflex was used.

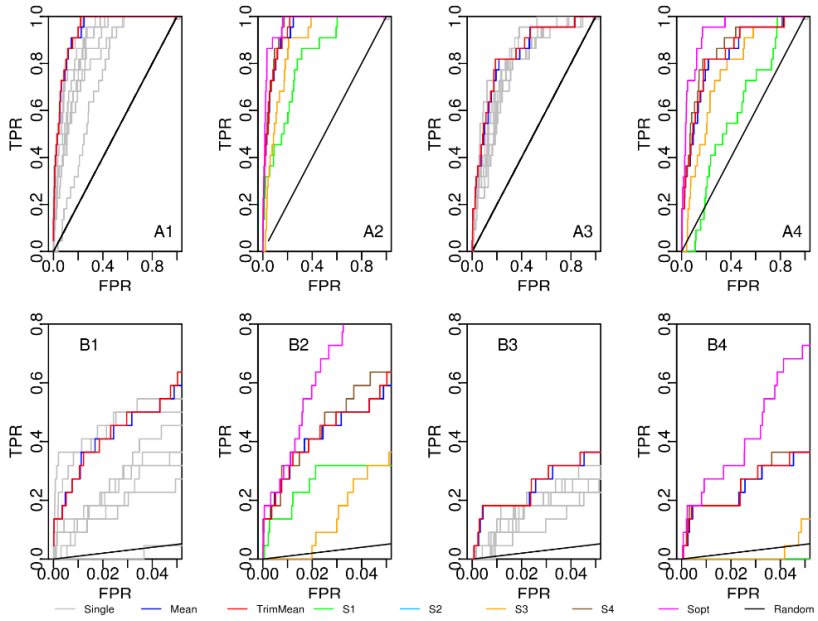


Figure 5. Receiver operating characteristic (ROC) curves for target Thymidine Kinase (*TK*). Fig. A1/A2: ROC curves by a rigid body docking program FRED for eight individual conformers (i) (gray) and seven consensus curves (S_{Mean} served as the benchmark in A2). **Fig. A3/A4:** analogous to A1/A2 these are ROC curves calculated after Surflex docking. **Fig. B1-B4**, the same data set and docking protocols as (A1-A4) focusing on the first 5% of the ROC curve. The black line (Random) represents random performance.

FRED and Surflex performance

Figure 5 describes the ROC curves for the eight individual conformers when either FRED (Figure 5A1,5A2) or Surflex (Figure 5A3,5A4) were used for target TK. FRED docking results (Figure

5A1) show that all of the active compounds were retrieved at 0.25 and 0.22 of the FPR by using S_{Mean} and $S_{TrimMean}$.

FRED steadily detected increasing numbers of active ligands in these conformers with nearly 80 to 90% of the total number of active ligands found within the top 0.2 to 0.6 of the FPR. From the ROC curve (Figure 5B1) it can be concluded that at the beginning, TPR derived from S_{Mean} and $S_{TrimMean}$ were rising steeply which indicates that the consensus methods S_{Mean} and $S_{TrimMean}$ overall performed better (p-value=0.07) than the other methods which were used with the eight individual conformers (Figure 5A1, Table 4).

Comparison to the enrichments of the other consensus methods (Figure 5A2) shows that S_{opt} yielded the best enrichment curve with an AUC value of 0.969 (Table 5), while S_{Mean} and $S_{TrimMean}$ were best at the early screening as demonstrated in Figure 5B2.

Figures 5A3 and 5A4 describe the ROC curves based on the Surflex docking results. The consensus methods S_{Mean} and $S_{TrimMean}$ gave the best enrichments and AUC values (Table 5) as compared to the eight individual conformers (Figure 5A3). As was the case for FRED, the consensus methods S_{opt}, S_{Mean} and $S_{TrimMean}$ yielded the best enrichments in the entire database (Table 4, Table 5). The methods S_2, S_4 were similar to the benchmark (S_{Mean}), whereas S_1 and S_3 performed worse than benchmark (Figure 5A4, Table 4, Table 5).

The ROC curves of the individual conformers and 'multiple' conformers for the targets: NA, FXa and TH have been calculated and analyzed as presented here for target TK. Overall, the optimal

performance of consensus methods S_{opt}, S_{Mean} and $S_{TrimMean}$ is observed also for these targets (Figures 6, 7, 8).

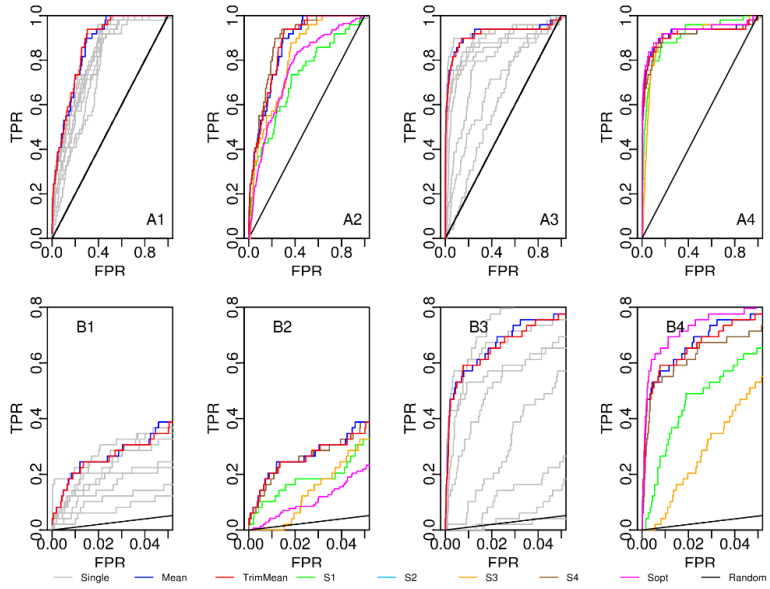


Figure 6. Receiver operating characteristic (ROC) curves for target Neuraminidase (*NA*). Fig. A1/A2: ROC curves by the rigid body docking program FRED for eight individual conformers (*i*) (gray) and seven consensus curves (S_{Mean} served as the benchmark in **A2**). **Fig. A3/A4**: ROC curves by Surflex. **Fig. B1-B4,** the same data set and docking protocol as (A1-A4) focusing on the first 5% of the ROC curve. The black line (Random) represents random performance.

Typically, the compounds at 1% of a ranked database are screened experimentally. Therefore besides the ROC curve, we chose to analyze additionally the various enrichment factors (EF) at 1% of the database for each of the eight conformers of the four protein targets and for S_{opt}, S_{Mean} and $S_{TrimMean}$ via FRED and Surflex. Generally, S_{opt}, S_{Mean} and $S_{TrimMean}$ resulted in the best or second best enrichment at 1% during different targets either FRED or Surflex (Table 6).

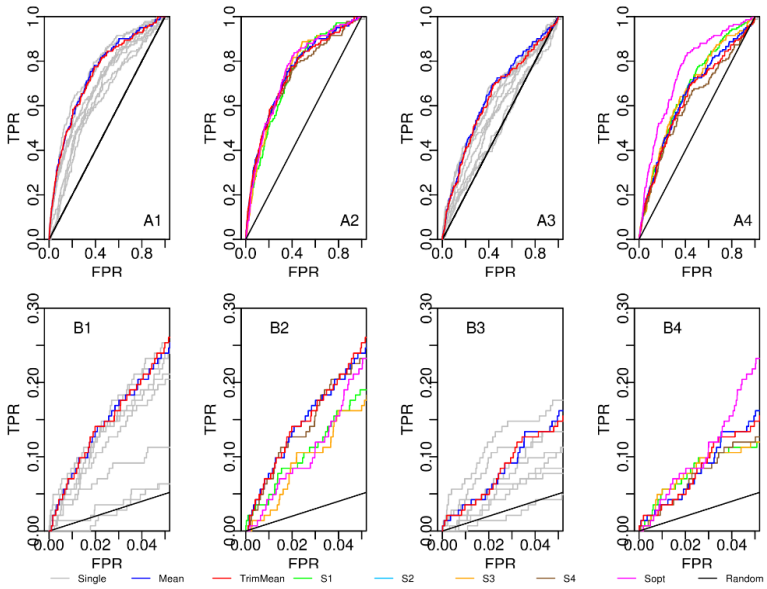


Figure 7. Receiver operating characteristic (ROC) curves for target *FXa*. Fig. A1/A2: ROC curves by the rigid body docking program FRED for eight individual conformers (*i*) (gray) and seven consensus curves (S_{Mean} served as the benchmark in **A2**). **Fig. A3/A4**: ROC curves by Surflex. **Fig. B1-B4,** the same data set and docking protocol as (A1-A4) focusing on the first 5% of the ROC curve. The black line (Random) represents random performance.

FRED and Surflex consensus performance

The FRED-Surflex combined ROC curves for the eight conformers of the four protein targets, TK, NA, FXa and TH, were calculated via two different methods: *sepcom* and *concom* (Figure 9, 10, 11, 12). With the method *concom*, eight conformers shared the same dataset after FRED and produced eight enrichment curves after Surflex (red curves in Figure 9). The *concom* method performed better than *sepcom* for the entire dataset in all conformers of TK (Figure 9).

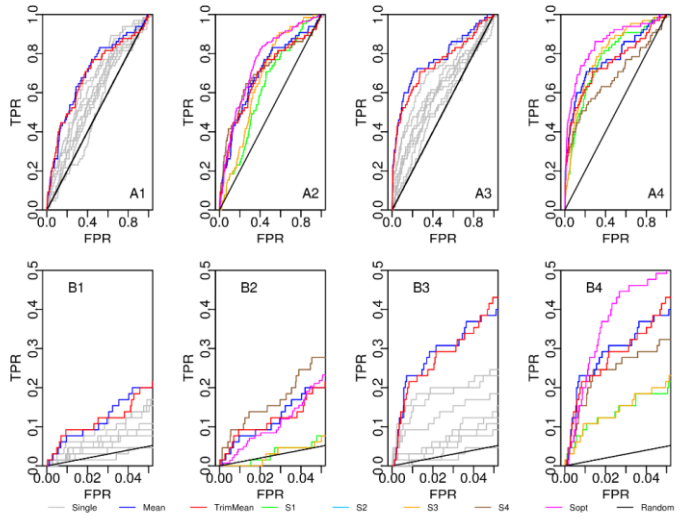


Figure 8. Receiver operating characteristic (ROC) curves for target Thrombin *(TH)*. Fig. A1/A2: ROC curves by the rigid body docking program FRED for eight individual conformers (*i*) (gray) and seven consensus curves (S_{Mean} served as the benchmark in **A2**). **Fig. A3/A4**: ROC curves by Surflex. **Fig. B1-B4,** the same data set and docking protocol as (A1-A4) focusing on the first 5% of the ROC curve. The black line (Random) represents random performance.

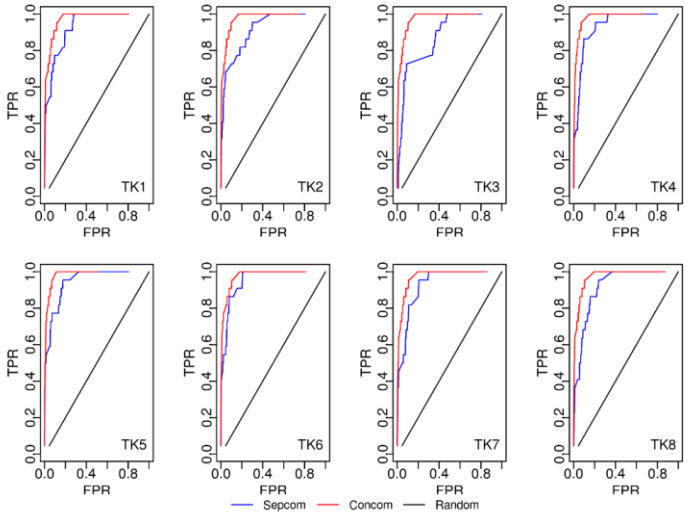


Figure 9. Comparison of ROC curves for *concom* (Red) and *sepcom* (Blue) consensus protocols for the target *TK*. X-axis: FPR: false positive rate as defined in de Methods; Y-axis: TPR: true positive rate as defined in the methods section. The dotted curve (Control) represents random performance. For *concom*, method $\boldsymbol{S_{Mean}}$ was employed here.

Similar analyses were performed for targets NA, FXa and TH (see Figures 10, 11, 12). For NA, overall, the ROC curves corresponding to the *concom* method were better than the *sepcom* curves except in conformer 2, where the *concom* is similar to *sepcom*. For target FXa the *concom* method performed better in six conformers, yet in the other two conformers, *concom* and *sepcom* perform similarly. Likewise, for target TH, the performance of *concom* was better than that of *sepcom* for all conformers throughout the entire database.

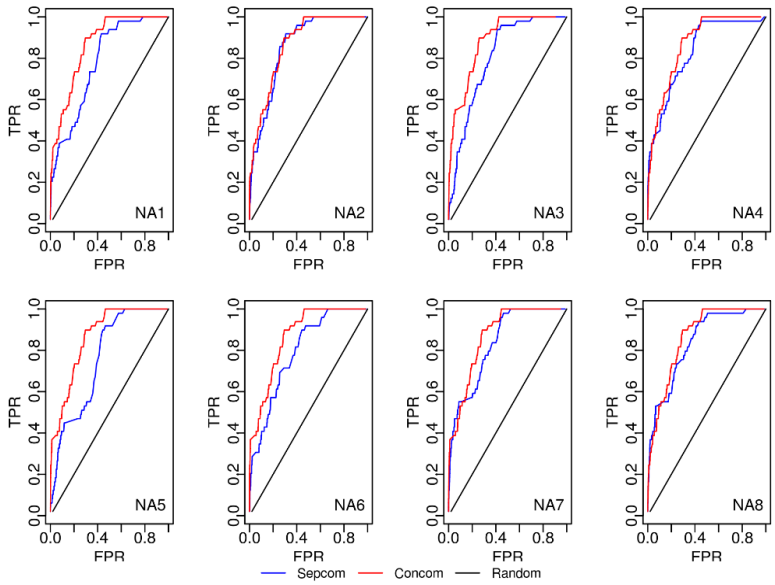


Figure 10. ROC curves for two different consensus protocols: *concom* (Red) and *sepcom* (Blue) in the target *NA*. TPR: True Positive Rate and FPR: False Positive Rate.

When the data for the individual target conformers were combined, using either methods S_{Mean}, S_1, S_2, S_3, S_4, or S_{opt}, we obtained the results as shown in Figure 13. In the ROC curves in Figure 13A1-A4, the S_{Mean} curve serves as benchmark. For TK, the methods S_{opt} and S_4 performed better than the benchmark, with S_{opt} being the best

(with around 0.85 of TPR at the top 1% of the database), while S_1 and S_3, performed worse than the benchmark (Figure 13B1 and Table 4). Notably, at 1% of the compound database, the combination of FRED and Surflex, while taking into account the results from different individual conformers (Figures 5 and 13A1,13B1) for target TK resulted in a net improvement of several folds as compared to FRED or Surflex only.

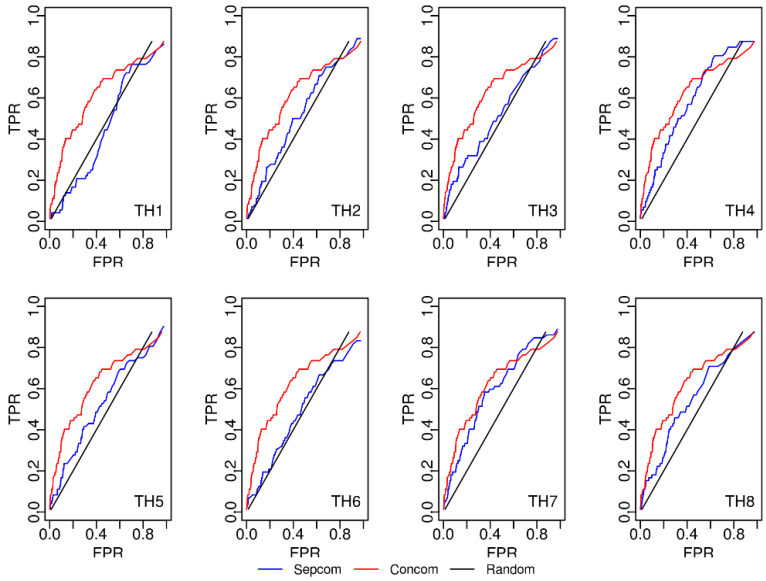


Figure 11. ROC curves for two different consensus protocols: *concom* (Red) and *sepcom* (Blue) in the target *FXa*. TPR: True Positive Rate and FPR: False Positive Rate.

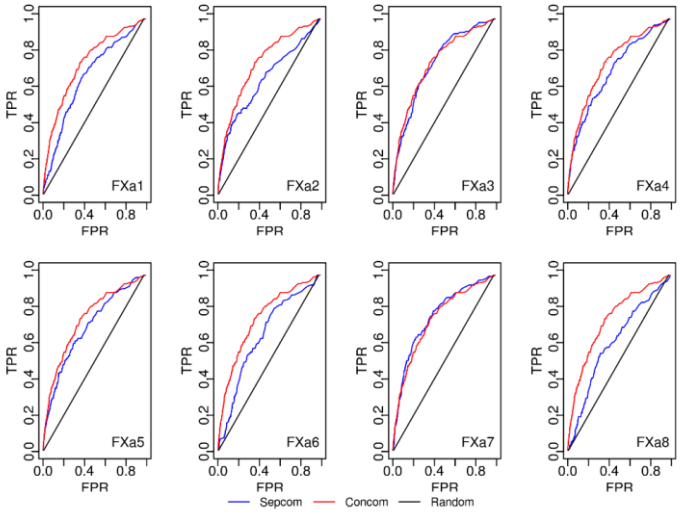


Figure 12. ROC curves for two different consensus protocols: *concom* (Red) and *sepcom* (Blue) in the target *TH*. TPR: True Positive Rate and FPR: False Positive Rate.

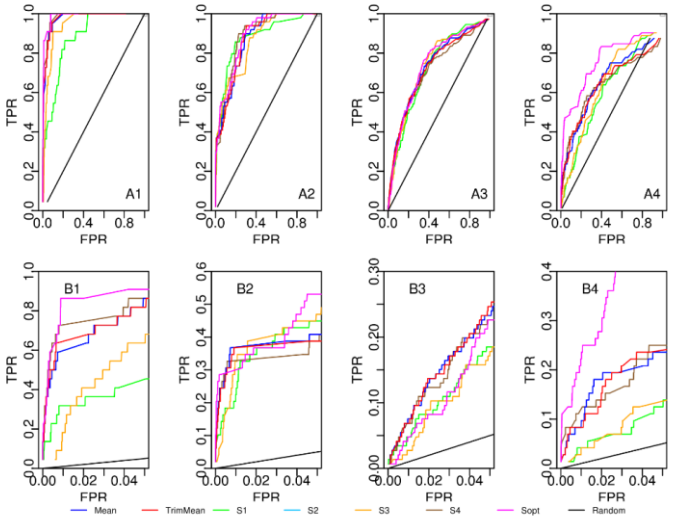


Figure 13. Receiver operating characteristic (ROC) curves by "*concom*" combined FRED-Surflex VLS approach for four targets by using the different consensus methods to screen the whole compound dataset. A1: *TK*, A2: *NA*, A3: *FXa*, A4: *TH*., and **Figure B1-B4**, the same data set and docking protocol as (A1-A4) focusing on the first 5% of the ROC curve. The black line represents random performance.

Figure 13A2 shows a similar analysis for target NA. Consensus methods S_2, S_4 and S_{opt} performed similar to the benchmark (Table 4, Table 5), in particular at the lowest percentages of the database (Figure 13B2). Method S_1 and S_3 performed worse than the benchmarks in 1% FPR fraction for this target (Figure 13B2, Table 4). The ROC curve for FXa is shown in Figure 13A3. All consensus methods performed comparable to the benchmark (p-value shown in Table 4). At 5% FPR fraction (Figure 13B3), S_1, S_3, and S_{opt} underperformed while S_2, S_4 were similar as the benchmark (S_{mean}). Overall, the method performance was comparable for all methods for this protein target. For TH (Figures 13A4 and 13B4), S_{opt} performed the best with the AUC value as 0.756. The TPR around 0.5 was retrieved at 4% of the dataset and up to 0.8 of the TPR at 40% of the database.

Stepwise improvements of combining FRED and Surflex

To obtain an insight into the improvement introduced in the overall method when applying Surflex subsequent to FRED-docking, and to study the improvements by variable top ranked lists of FRED being subjected to Surflex docking, we calculated stepwise improvement curves. The stepwise improvement curves showed the enrichment improvements per percentage step of the database when both FRED and Surflex were combined via *concom*. This was calculated for the entire database and for each of the four protein targets (Figure 14). The results are most obvious for target TK and show that for the majority of the curves (S_{Mean}/$S_{TrimMean}$, S_1, S_2, S_3, S_4, and S_{opt}) calculated (Figure 14A), a high frequency of improvement was

reached within the first 20-30% of the database, with the improvement (Z) ranging from 50 to 100%. Notably S_{opt} resulted in lower frequencies of improvement as compared to the other curves. Some curves exhibited negative Z values with the exception of S_1, S_3, and S_{opt}. The overall trend in frequencies is from high (within the top 10% of the database) to low (10~30% of the database) and zero (beyond 30% of the database) with the exception of S_1 and S_3, which indicates that in particular in the lower percentages of compound database tested, the combination of FRED and Surflex is beneficial. Figure 14B shows the enrichment improvements for target NA. Independent of the consensus method applied, Surflex increased the enrichments, as judged from overall positive Z values calculated. All curves in Figure 14B show a high frequency of improvement within the first 25~45% of the database with improvements (Z) ranging from 50 to 100%. Only S_1 appeared to be distributed over the complete range of percentages tested, albeit with highest frequency concentrations between 0% and 40% of the database. Some curves have negative Z values, representing an overall lower quality at a particular percentage of the database due to the combination of the rigid and flexible methods, with S_4 having the most.

Figure 14C shows the enrichment improvement graphs for FXa. For all methods analyzed here, inclusion of Surflex in the protocol overall increased the enrichments. For this protein, the majority of curves showed a high frequency of improvement throughout the entire database with improvements (Z). Remarkably, negative Z values were calculated exclusively between 0 and 40% of the database. Enrichment improvement graphs for target TH are shown in Figure

14D. Like in target FXa, Surflex appeared to increase the enrichment for the entire set of methods tested. All curves in target TH (Figure 14D) show high frequencies of improvement throughout the entire database with the improvement (Z) ranging from 50 to 100%. Negative Z values appeared mostly at the beginning of the database (0-40%).

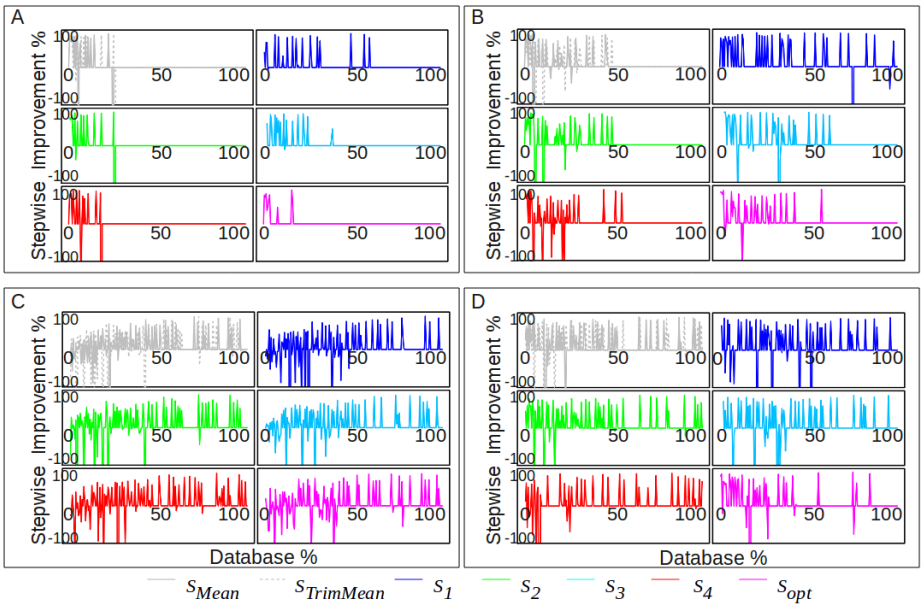


Figure 14. Stepwise improvement by combining Surflex to FRED for four targets by seven different consensus methods. A positive value at *X* % means that the overall combined FRED-Surflex procedure is improved whereas a negative value indicates that FRED only performs better than that from the combined procedure at the region between *X* % to (*X+1)* %. X-axis indicates percentage of the total database (0-100%), while Y-axis shows the degree of improvement. **A:** *TK*, **B:** *NA*, **C:** *FXa*, **D:** *TH*, *concom* was used to combine ranked lists.

Screening of directory of useful decoys (DUD) data sets

Although the new protocol was tested on four diverse targets: TK, NA, FXa, TH we also applied the protocol to 6 other targets: GART, CDK2, Trypsin, COMT, ER and INHA to ascertain the wider

applicability of the method. The compound datasets (active and decoys) applied for virtual screening for these 10 targets were obtained from the DUD dataset. The number of actives / decoys and the diversity of those actives are shown in Table 3 (64).

First, the performance of the docking protocol by using either single or multiple conformations was compared. The use of 8 different conformers, as was determined from data shown in Figure 1, appears valid given that AUC do not drop by inclusion of eight conformers, instead for targets thrombin, ER and 1NHA, use of eight structures appears optimal with the given data sets from the DUD for these targets (Figure 2). FRED docking yielded comparable the area under the ROC curve (AUC) values (Table 5) for both single and multiple conformations in some cases, such as TK, Thrombin, COMT and INHA. However, by inspecting the entire ROC curves for all targets as shown in (Figure 15, 16, 17, 18), multiple conformations performed better than using single conformation for FRED docking. On the other hand, AUC values (Table 5) derived from the single conformation using Surflex docking are mostly lower (p-value <0.05; Table 7; Figure 19, 20, 21, 22) than those obtained from applying multiple conformations for docking, indicating that the use of multiple conformations can result in higher enrichment than use of single conformation for Surflex docking.

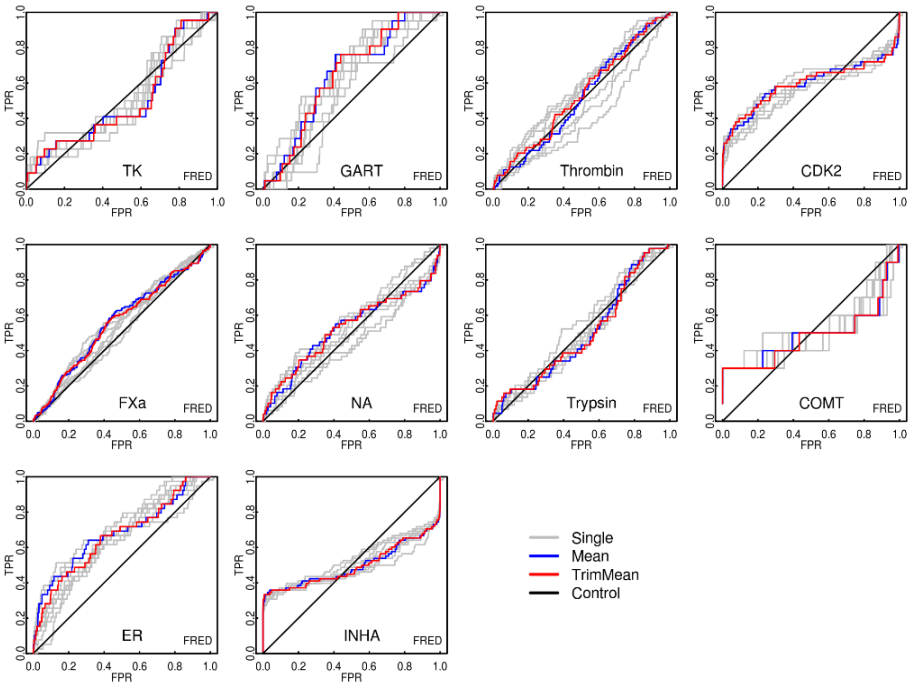


Figure 15. Receiver operating characteristic (ROC) curves to compare multiple conformers (Blue, Red) and single conformers (gray). 10 Targets were tested with their corresponding DUD dataset by ***FRED*** docking. Each target contains eight conformers, which are shown as gray ROC curves. Calculated *Mean* (Blue) and *TrimMean* (Red) are shown to indicate the multiple conformations. TPR: True Positive Rate and FPR: False Positive Rate.

Next, the performance of the docking using either FRED or Surflex alone was compared with the combined FRED-Surflex method. The AUC (Table 5) and ROC curves (Figure 23, 24) of these 10 targets derived from the DUD data sets show that the combined FRED/Surflex methods performed better than using FRED docking only, whereas the combined FRED-Surflex methods gave comparable results to the Surflex docking (p-value >0.05). However, docking by using the combined method is much faster than using Surflex docking alone because some false positive compounds were first filtered out by using the fast FRED docking protocol and in a

typical virtual ligand screening campaign, the dataset being used for Surflex docking can be geometrically filtered by FRED. Therefore, in order to speed up calculations and save computational costs, the combined FRED-Surflex method is suggested to be applied for screening large compound databases. Finally the different consensus scorings (S_{Mean}, $S_{TrimMean}$, S_1, S_2, S_3, S_4, and S_{opt}) were compared. Results as demonstrated by the AUC, ROC curves and p-value indicate that S_{opt} outperformed the other consensus scoring methods.

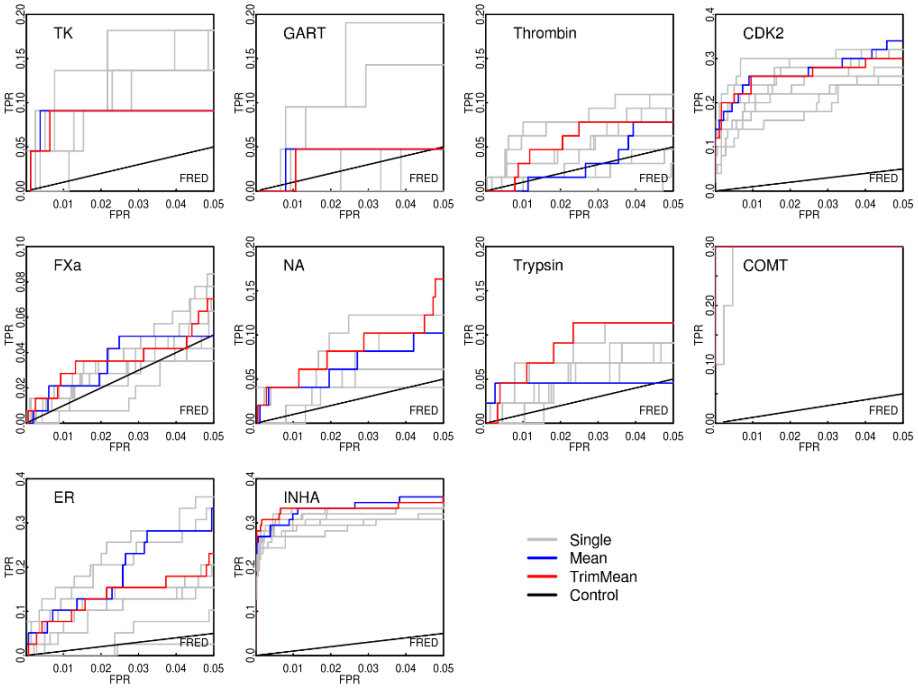


Figure 16. Receiver operating characteristic (ROC) curves at the 5% false positive fraction to compare the multiple conformers (Blue, Red) and single conformer (gray). 10 Targets were tested with their corresponding DUD dataset by ***FRED*** docking as in Figure 15. Each target contains eight conformers, which are shown as gray ROC curves. Calculated *Mean* (Blue) and *TrimMean* (Red) are shown to indicate the multiple conformations. TPR: True Positive Rate and FPR: False Positive Rate.

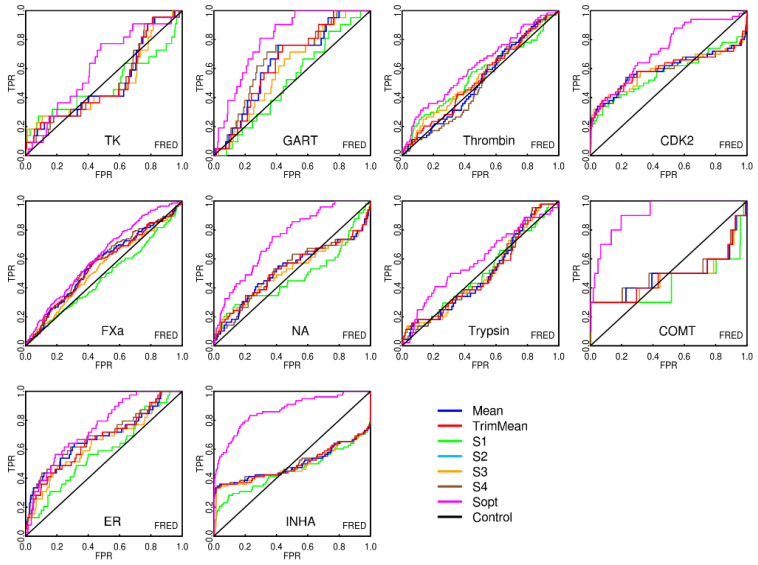


Figure 17. Receiver operating characteristic (ROC) curves to compare 7 consensus methods. 10 Targets were tested with their corresponding DUD dataset by ***FRED*** docking. S_{Mean} and $S_{TrimMean}$ served as the benchmark TPR: True Positive Rate and FPR: False Positive Rate.

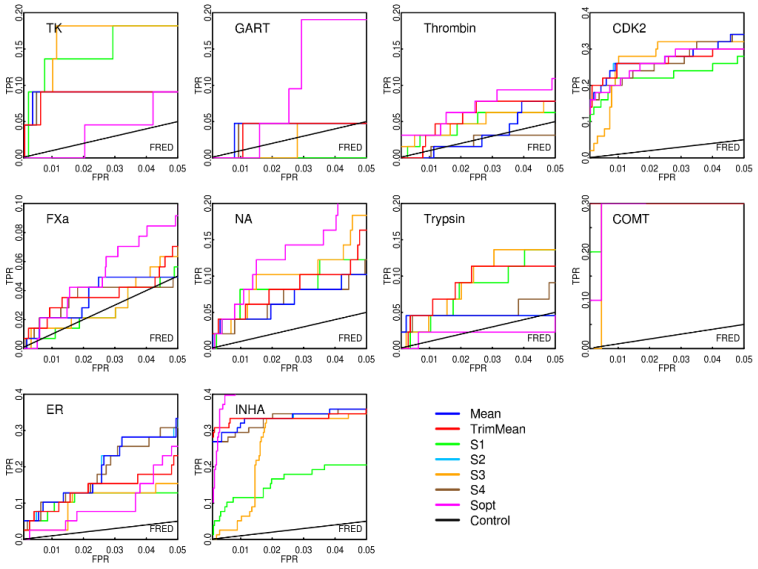


Figure 18. Receiver operating characteristic (ROC) curves at the 5% false positive fraction to compare 7 consensus methods.

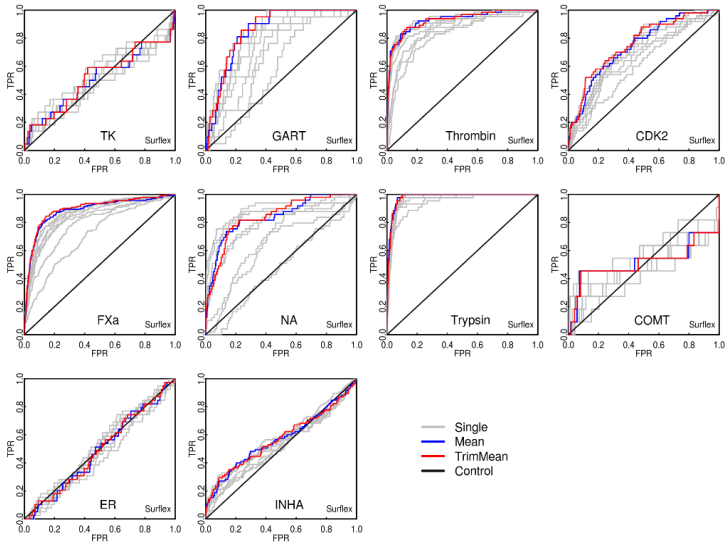


Figure 19. Receiver operating characteristic (ROC) curves to compare the multiple conformers (Blue, Red) and single conformer (gray). 10 Targets were tested with their corresponding DUD dataset by ***Surflex*** docking. Each target contains eight conformers which are shown in gray. Calculated Mean (Blue) and TrimMean (Red) are shown to indicate the multiple conformations. TPR: True Positive Rate and FPR: False Positive Rate.

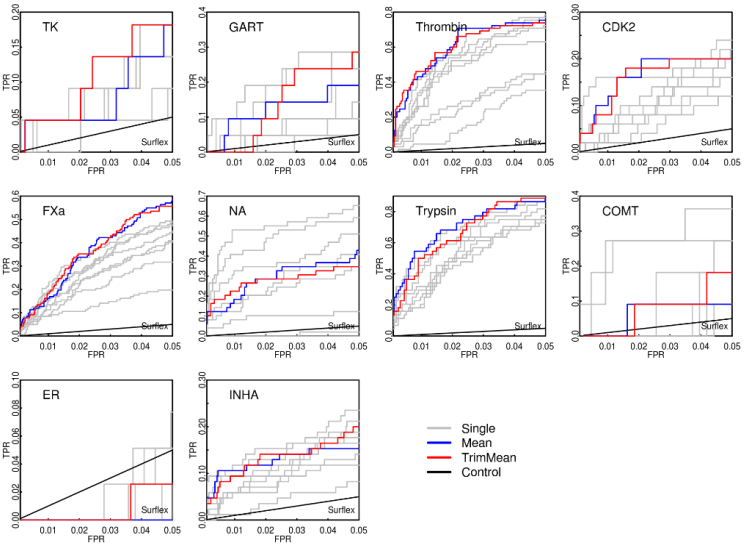


Figure 20. Receiver operating characteristic (ROC) curves at the 5% false positive fraction to compare the multiple conformers (Blue, Red) and single conformer (gray).

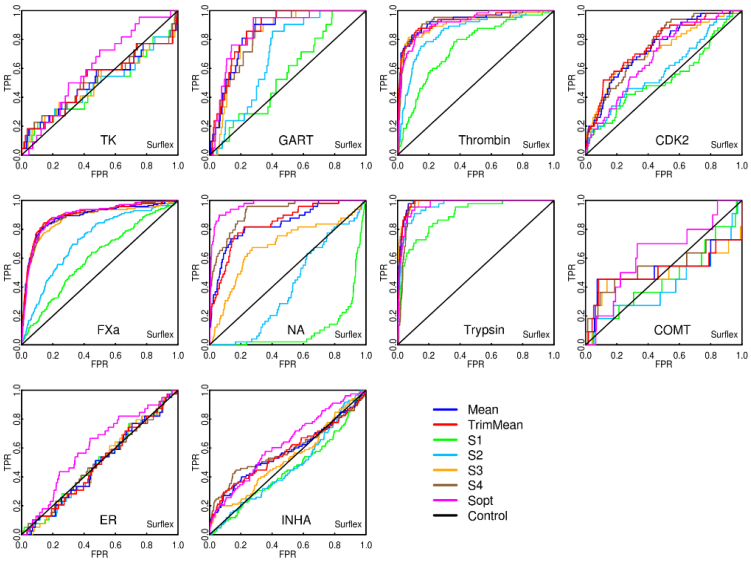


Figure 21. Receiver operating characteristic (ROC) curves to compare 7 consensus methods. 10 Targets were tested with their corresponding DUD dataset by ***Surflex*** docking. S_{Mean} and $S_{TrimMean}$ served as the benchmark TPR: True Positive Rate and FPR: False Positive Rate.

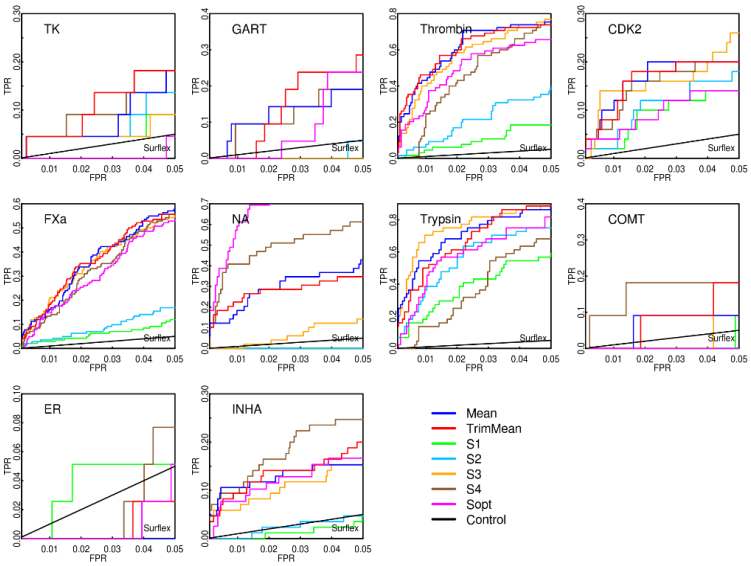


Figure 22. Receiver operating characteristic (ROC) curves at the 5% false positive fraction to compare 7 consensus methods.

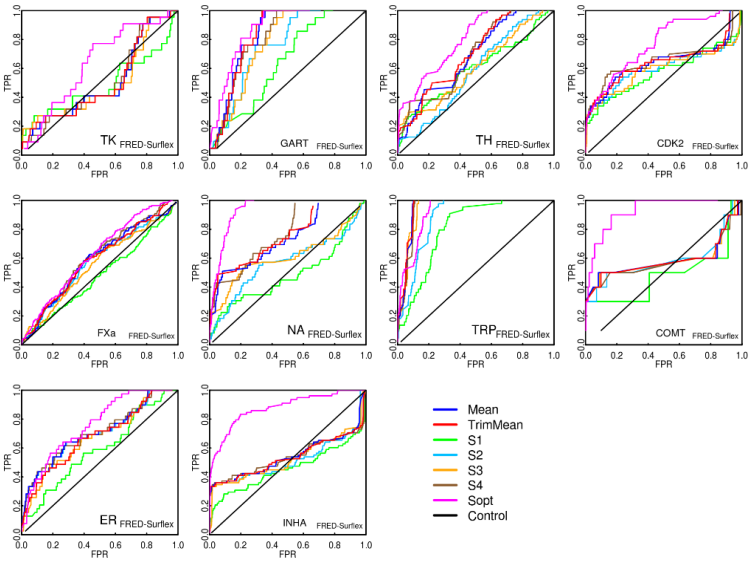


Figure 23. Receiver operating characteristic (ROC) curves to compare 7 consensus methods. 10 Targets were tested with their corresponding DUD dataset by **combined FRED-*Surflex*** docking. S_{Mean} and $S_{TrimMean}$ served as the benchmark TPR: True Positive Rate and FPR: False Positive Rate.

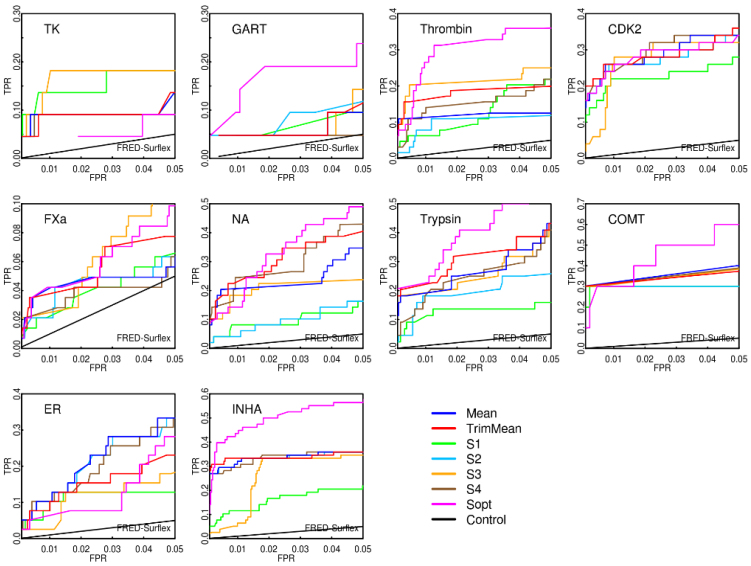


Figure 24. Receiver operating characteristic (ROC) curves at the 5% false positive fraction to compare 7 consensus methods.

Discussion

In VLS, one wishes to achieve an optimal trade-off between highest possible compound specificity, target-affinity, time requirement of in silico hit identification and a lowest possible effort for actual in vitro activity testing. This implies that a compromise will have to be reached to obtain an optimal overall method for hit identification. In this study we particularly explored the use of ensemble docking and how to combine results from multiple target conformers into a consensus ranked list for an existing hierarchic VLS method that combines two docking approaches, FRED and Surflex.

The rigid body docking software FRED was used to rapidly screen a test compound database of 59,479 molecules. Essentially the use of rigid body docking software works as a first filter, comparable to a pharmacophore filter, to reduce the number of potential molecules for flexible docking. With the current progress in flexible docking, inclusion of rigid body docking may be disputed, however, especially in the absence of known ligands (e.g. for new targets) use of such rigid body docking filters may prove beneficial to overall method efficiency.

As might be expected, the CPU time of FRED appeared to correlate with the size of the protein target's pocket, except for target NA (Table 2). Surflex running times did not correlate with the pocket sizes of the conformer targets (Table 2). This represents the fact that in case of flexible docking not only pocket size, but also other pocket properties such as shape and hydrogen bonds are factors that

influence the Surflex docking time (24).

To evaluate the combined use of FRED and Surflex, we first calculated enrichment performances of FRED and Surflex separately. For target TK (Figures 5A1-A4 and 5B1-B4), FRED yielded better enrichments than Surflex. The enrichment by Surflex only is better than by FRED only for target NA (Figure 6) and to a lesser extent for TH (Figure 8). For target FXa (Figure 7) the ROC curves were similar for Surflex and FRED. From these diverse results, we conclude that flexible docking programs are not by definition better than rigid ones in VLS and that the properties of the target site have a major influence on this (17). Chance ranked list outliers do not appear to influence overall enrichments considerably, since the method S_{Mean} performed always better in ROC curves than method $S_{TrimMean}$. At the practically important early enrichment of ROC curve (typically 0-5% of the ranked database) the averaging of enrichments over the eight conformers would yield a priori higher percentages of actives than if any of the single conformers would randomly have been chosen. Few individual curves, both for FRED and Surflex, presented percentages higher than S_{Mean} or $S_{TrimMean}$ (see e.g. Figures 5A1 and 5A3, 6A1, 6A3, 7A1, 7A3, 8A1 and 8A3), however in the absence of prior knowledge about the correct target conformation it is impossible to select the conformation which will yield the highest number of actives to be identified (48). Interestingly, the S_{Mean}, $S_{TrimMean}$ processed data for target TH yielded comparatively much better enrichments than that in any individual conformer (Figures 8A1 and 8A3). Since for all targets the consensus curve is overall better than the ones from single conformations, multiple conformations should

be selected in order to perform a SBVLS as reliably as possible. This may not only reflect a biological relevance (65, 66), but is also based on statistical improvement of the enrichment of a screening program.

After having established that use of multiple conformations of a target is optimal for the current VLS protocol, the selection of an optimal consensus method to integrate the information from the different conformations is a crucial factor for overall docking quality. We have used seven methods for the calculation of a consensus score. These methods are based on the docking scores for each compound in different conformers, taking into account the standard deviation (σ) of scoring which is imposed into the methods. The different consensus methods studied yield different ROC curves. S_{Mean}, S_2, S_4, and S_{opt} have improved enrichments while S_1 and S_3 show only a modest influence on the enrichment. As compared to the benchmark (S_{Mean}), the method S_{opt} performed best, followed by method S_4 for all the four targets, whereas the methods S_1 and S_3 do not result in improvement of ROC curves over those of the benchmark (Figures 5A2 and 5A4, 6A2, 6A4, 7A2, 7A4, 8A2, 8A4). Therefore methods S_1 and S_3 are not further included in the discussion.

For all targets used in our study, an increasing interval between two active ligands (represented by peak interval in Figure 14) indicates an improvement of the FRED-processed database by subsequent Surflex docking. The improvement as a function of database size, as shown in Figure 14, is an important parameter for cut-off selection. For example, in target TK a meaningful cut-off would be the top 18% of FRED results (S_{opt} and S_4) or the top 20% of FRED results (S_2).

For target NA the cutoff would be 25% (S_4), 42% (S_{opt}, S_3) and 50% (S_2). For targets NA and TH (Figures 14B and 14D), use of S_{opt} appears optimal with cut-offs at ~43% of the total database. Overall, inclusion of Surflex in the docking protocol improved the enrichments at most but not all of the steps. However, since the number of positive peaks is more than the number of negative peaks, Surflex improves the enrichment of the combined FRED-Surflex process as a whole. Interestingly, at some steps Surflex worsened the enrichment curves. We cannot exclude that some compounds in our dataset are false negatives, which in fact represent unknown active ligands.

When multiple conformers are chosen for rigid/flexible docking, two methods can be applied to select compounds from FRED results for consecutive Surflex docking as described in the methods section (see also Figure 4). The enrichments of *concom* are better than the ones of *sepcom* in all targets TK, NA, FXa and TH. Therefore we propose that in particular for the proteins tested here, as a model for the class they represent, it is advisable to use the consensus combination method (*concom*) for the combined FRED-Surflex VLS approach.

The combined FRED-Surflex enrichments have been calculated while employing several consensus methods. The combined FRED-Surflex results for targets TK, NA, FXa and TH (Figures 13A1, 13A2, 13A3 and 13A4 respectively) are better than FRED only results (Figures 5A2, 6A2, 7A2, 8A2). For targets TK, FXa and TH, but not NA the combined results are better than Surflex only results (Figures

5A4, 6A4, 7A4, 8A4). The considerable improvement in ROC curve that is achieved by combination of conformers in case of Surflex makes it questionable to include FRED in the protocol for NA. Likewise, the improvement by averaging multiple conformers rank lists for this particular target is more evident than for the other protein targets. In the case of FXa, the combination of FRED and Surflex did not result in clear enrichment improvements, while neither FRED nor Surflex showed good enrichment for this target. This would suggest that a FRED-Surflex combination cannot result in an improved enrichment curve as compared to FRED or Surflex only if a first FRED docking step does not result in a reasonable enrichment during a first VLS step (Figures 13C and 7). By using our data sets (active from DUD combined with decoys from ChemBridge) we found that inclusion of FRED as a first filter is however beneficial in 3 out of 4 targets tested here and by itself computationally not intensive, whereas it allows the exclusion of part of the small molecule database that need not be tested via the computationally more intensive flexible method.

The number of positive peaks exceeds that of the negative peaks in the stepwise improvement performance (Figure 14) which indicates that the enrichment of the combined FRED-Surflex is better than that of either FRED or Surflex alone. In terms of the stepwise improvement by easily calculatable S_{Mean}, the efficient cut-off of the ranked FRED results may be used as input for Surflex, with cut-offs of 25% in TK, 47% in NA, 61% in FXa, 52% in TH. For the better performing S_{opt} method these percentages are 17%, 42%, 47% and 43% respectively. We conclude that ideally the cut-off for any type of

target (with its characteristic pocket properties) is determined, however, in the absence of information on an individual target, an average 46% (using S_{Mean}) of the ranked FRED results may be used as an input for Surflex screening.

Finally, we have tested our virtual screening protocol by using actives and decoys derived from the standard DUD data sets for 10 targets (TK, NA, FXa, TH, GART, CDK2, Trypsin, COMT, ER and INHA). In general, results obtained from these DUD data sets are similar to the results which were obtained from the 4 targets (TK, NA, FXa, and TH) using decoys from Chembridge database. Multiple conformations showed better performance than using single conformation (see Figures 15, 16, 19, 20). Therefore in the absence of information about which protein conformation should be used in docking, the multiple conformations method is suggested. The combined FRED/Surflex docking performed better than using FRED or Surflex docking alone as indicated by the AUC and ROC curves (compare Figures 17, 18, 21, 22 and 23, 24 and Table 5). However, the performance of the combined methods depends on the docking results of FRED as discussed above. In this study we used general standard FRED docking parameters for all target proteins while it is possible to first optimize and validate parameters in FRED docking for a specific target protein which can help to improve the enrichment of screening. When comparing consensus scoring methods, we can summarize that S_{opt} apparently outperforms the other consensus scoring methods, and S_4 is better than S_{Mean} which is the benchmark for evaluating the consensus methods. Hence, in the case where no

active compounds are available for generation of S_{opt}, use of S_4 consensus scoring is advisable.

Conclusion

We have screened a database of 59,479 compounds by a novel method, that uses a combination of the programs FRED and Surflex in a parallel computational environment, taking into account results from multiple target conformations. In this study, four different targets (TK, NA, FXa and TH) with each eight conformations were selected. Although the enrichment results varied for the different targets, reflecting the chemical properties and size of these different target pockets, they indicate that the use of multiple conformers for a given target is preferable, while we realize that the chosen four targets do not fully represent the wide variability that exists with respect to targetable pockets. To optimally combine the results of the eight different conformers, it is advisable to apply a consensus of the FRED results for the different conformers and from that generate a sub dataset for consecutive Surflex screening, (so-called "*concom*" method in this paper). The cut-off of the ranked FRED results which are further applied as input for Surflex is a key step for a successful virtual screening. For the targets chosen here, no general standard cut-off value can be given, and this value must be first investigated for each target protein. We have demonstrated however that the stepwise improvement is a useful method to detect the optimal cutoff of a FRED docked list, for further processing by Surflex docking.

We also applied our virtual screening protocol on 10 different targets

(TK, NA, FXa, TH, GART, CDK2, Trypsin, COMT, ER and INHA) using active and decoys from the DUD data sets. Results obtained from these data sets are in agreement with our findings for the 4 targets discussed before.

The quality of FRED enrichments appears crucial to achieve an improved result for the combined FRED-Surflex VLS approach. Therefore, before using the combined method, one may consider to design an appropriate FRED screening protocol. In this study, we present a new consensus method, S_{opt}, based on the results of an optimization problem. This method is designed to produce consensus lists for a protein target with an optimal enrichment for both FRED and Surflex but requires prior knowledge about confirmed actives. In the absence of such knowledge, use of the easily accessible methods S_{Mean} and S_4 is advisable.

Table 1. Pocket information of protein conformers (4 * 8). The mutable residues heavily influence the docking accuracy and are potential hydrogen donors /acceptors. The structural diversity of the conformers is indicated by their pocket size, outer and inner contours which are used for FRED docking and served as references for further Surflex docking are listed.

Conformer	Mutable Residues	Volume of Box ($Å^3$)	Outer Contour ($Å^3$)	Inner Contour ($Å^3$)
TK1	H14, T1:T22, S30, Y36, T42, Y43	9271	1960	33
TK2	H14, T19:T22, Y36, T42:Y43, Y57, Y88	7726	1914	43
TK3	H14, T19:T22	4646	1014	91
TK4	Y9, H14, T21:T22	7591	1761	48
TK5	Y9, H14, T19, T21:T22, Y36, T42:Y43, Y57	6355	1609	51
TK6	H14, T19:T22, Y36, T2b:Y3b, Y17b, H28c, Y50e, S53e	6959	1535	55
TK7	Y7b, H12b, T17b, T19:T20b, Y34b, T40:Y41b, Y55b, Y86b, Y61c	5288	1517	61
TK8	H12b, T18:T20b, Y34b, T40:Y41b, Y86b, Y19c	4991	1556	72
NA1	N3b, Y118b	2400	1292	91
NA2	T34b, T81b, Y118:S119b	3888	1574	54
NA3	N3b, T34b, T81b, Y118:S119b, S40e	4418	1538	74
NA4	N3e, T34e, T13f, Y50:S51f	4116	1477	85
NA5	N3c, T34c, T81c, Y118:S119c, H20d, T133e	5039	1628	70
NA6	N3e, Y22j	2589	1332	88
NA7	N41c, N3e, Y76f, H20g	3394	1503	80

NA8	Y50j	2089	1194	88
FXa1	N20, H42, Y19b, S119b, Y7c	4347	1602	72
FXa2	H42, T18:Y19c, S49d, S68d, Y3e	2059	979	99
FXa3	H42, Y19b, S3e, S22e, Y33e	3678	1535	75
FXa4	Y19b, H9e, Y12e, Y7h	4128	1574	73
FXa5	H20b, T18:Y19d, S8g, Y3h	3334	1183	90
FXa6	H42, Y6e, H11f	1906	936	82
FXa7	H42, Y45, T18:Y19b, S11g, Y22g	3966	1634	68
FXa8	T12c, T25e, N27e	1717	1034	50
TH1	Y41b, S22:T23e, Y8f, T12f, H17h, T21h	4616	1680	35
TH2	N4b, Y36b, H38b	2983	1315	95
TH3	Y41:N42b, N46b, H17f, T12g, N14g	4986	1518	66
TH4	H43, Y47, Y10e, S7g	4391	1448	86
TH5	Y37b, N1e, T23g, T28g	2406	1206	99
TH6	H8b, Y12b, T2i, S46i	3083	1404	83
TH7	H8e, Y11:N12c, N1d, T6e, N2g, S15i, Y8j, Y11:H13j, T4l, H7l, Y11c	10524	1741	69
TH8	H6d, Y10d, N1f, T23h, T4i, N6i, S48i, T63i, N1j	7189	1769	50

Table 2. Pocket characteristics and CPU time consumption for the four protein targets.

Target ID	FRED	Surflex*	-fold difference	Pocket RMSD Range	Avg Pocket size (Å3)	Avg box size ($Å^3$)	S.C. FRED	S.C. Surflex
TK	19.5±4.05	302.0±10.7	15.5	0.32-0.52	396	6603	r = 0.52	N.S.
NA	19.8±4.00	224.8±13.0	11.4	0.53-1.17	321	3491	r = 0.18	N.S
FXa	15.4±3.88	228.5±13.5	14.9	1.02-1.96	362	3141	r = 0.97	N.S.
TH	20.8±5.53	245.1±17.5	11.8	1.25-2.65	443	5022	r = 0.79	N.S.

Givens are times in hours±1SD, n= 8; the –fold difference indicates the ratio Surflex CPU time/ FRED CPU time; The pocket heavy atom RMSDs for eight conformers ranges are shown in column 5.; S.C. indicates size-time correlation as indicated by the Pearson coefficient (r) with 0.0<r<0.09 indicating no correlation; 0.1<r<0.3 weak correlation; 0.3<r<0.5 moderate correlation and 0.5<r<1.0 indicating strong correlation; N.S. not significant

* Sequential time on the cluster is indicated here.

Table 3. The number of DUD ligands and decoys used in this study. The ligands for each target have been clustered by their physico-chemical diversity (67).

Target	Total ligands	Total decoys	Clusters of Ligands
TK	22	891	7
GART	40	879	5
Thrombin	72	2456	14
cdk2	72	2074	32
FXa	146	5745	19
NA	49	1874	7
Trypsin	49	1664	7
COMT	11	468	2
ER	39	1448	8
INHA	86	3266	23

Table 4. *p*-Value matrix for 7 consensus methods by calculation of ROC curves and AUC values for 4 targets with the Chembridge decoy data set after FRED or Surflex docking. 'Single' was calculated as the median of 8 individual ROC curves. Values below 0.1 are in bold to emphasize significance at the 90% level.

		Mean	TrimMean	S1	S2	S3	S4	Sopt
TK FRED	Single	**0.057**	**0.018**	0.551	**0.016**	0.751	**0.090**	**0.005**
	Mean		0.835	**0.042**	0.864	0.316	0.860	0.705
	TrimMean			**0.048**	0.964	0.340	0.778	0.664
	S1				**0.046**	0.432	**0.038**	**0.028**
	S2					0.330	0.798	0.677
	S3						0.298	0.254
	S4							0.753
NA FRED	Single	**0.024**	**0.020**	0.843	**0.024**	0.145	**0.011**	0.462
	Mean		0.944	**0.040**	0.989	0.428	0.777	0.130
	TrimMean			**0.034**	0.955	0.388	0.832	0.113
	S1				**0.039**	0.208	**0.019**	0.591
	S2					0.420	0.788	0.127
	S3						0.282	0.471
	S4							**0.072**
FXa FRED	Single	0.104	0.110	0.232	0.110	0.148	0.186	**0.074**
	Mean		0.979	0.666	0.978	0.859	0.762	0.874
	TrimMean			0.685	0.999	0.880	0.782	0.853
	S1				0.685	0.799	0.897	0.555
	S2					0.880	0.782	0.853
	S3						0.900	0.737
	S4							0.644
TH FRED	Single	**0.066**	0.199	0.957	0.199	0.671	**0.035**	**0.003**
	Mean		0.582	**0.075**	0.581	0.159	0.786	0.243
	TrimMean			0.219	0.999	0.391	0.410	**0.086**
	S1				0.219	0.710	0.040	**0.003**
	S2					0.391	0.410	0.086
	S3						**0.093**	**0.010**
	S4							0.371
TK Surflex	Single	0.200	0.175	0.644	0.177	0.469	0.133	**0.042**
	Mean		0.939	**0.081**	0.944	0.577	0.826	0.452
	TrimMean			**0.068**	0.995	0.526	0.886	0.499
	S1				**0.069**	0.236	**0.049**	**0.013**
	S2					0.530	0.881	0.496
	S3						0.437	0.190
	S4							0.595
	Single	**0.010**	**0.012**	**0.018**	**0.009**	**0.035**	**0.020**	**0.005**

NA Surflex	Mean		0.946	0.823	0.982	0.623	0.788	0.806
	TrimMean			0.876	0.929	0.671	0.840	0.754
	S1				0.805	0.788	0.964	0.638
	S2					0.607	0.771	0.823
	S3						0.823	0.460
	S4							0.607
FXa Surflex	Single	0.340	0.393	0.333	0.393	0.331	0.693	**0.025**
	Mean		0.921	0.989	0.921	0.986	0.576	0.199
	TrimMean			0.910	0.921	0.906	0.645	0.166
	S1				0.910	0.997	0.566	0.204
	S2					0.906	0.646	0.167
	S3						0.564	0.206
	S4							**0.065**
TH Surflex	Single	**0.031**	**0.070**	**0.041**	**0.031**	**0.014**	0.408	**0.002**
	Mean		0.732	0.913	0.995	0.771	0.186	0.317
	TrimMean			0.816	0.727	0.527	0.326	0.179
	S1				0.907	0.689	0.225	0.267
	S2					0.776	0.183	0.320
	S3						0.106	0.477
	S4							**0.020**
TK Combined	Mean		0.654	**0.028**	0.971	0.195	0.540	0.182
	TrimMean			**0.031**	0.908	0.225	0.447	0.159
	S1				**0.025**	**0.076**	**0.022**	**0.010**
	S2					0.171	0.606	0.202
	S3						0.152	**0.063**
	S4							0.236
NA Combined	Mean		0.473	0.984	0.966	0.874	0.504	0.739
	TrimMean			0.460	0.447	0.576	0.165	0.293
	S1				0.982	0.858	0.517	0.754
	S2					0.840	0.531	0.771
	S3						0.408	0.622
	S4							0.737
FXa Combined	Mean		0.921	0.989	0.920	0.985	0.576	0.403
	TrimMean			0.909	0.999	0.906	0.645	0.349
	S1				0.909	0.996	0.566	0.411
	S2					0.906	0.645	0.349
	S3						0.563	0.413
	S4							0.163
TH Combined	Mean		0.858	0.112	0.899	0.638	0.982	**0.067**
	TrimMean			0.159	0.958	0.771	0.875	**0.045**
	S1				0.144	0.263	0.117	**0.001**
	S2					0.731	0.917	**0.051**

	S3						0.654	**0.021**
	S4							**0.064**

Table 5. AUC Metrics of different methods by FRED, Surflex, and combined FRED-Surflex for 10 targets. The highest value and S_{opt} value are shown in bold.

Chembridge decoys		Single	Mean	Trim Mean	S1	S2	S3	S4	Sopt
TK	FRED	0.865	0.940	0.943	0.817	0.940	0.880	0.951	0.969
	Surflex	0.803	0.835	0.843	0.594	0.835	0.763	0.862	0.937
	Combined		0.970	0.944	0.840	0.957	0.932	0.905	0.906
NA	FRED	0.799	0.859	0.864	0.714	0.860	0.803	0.879	0.752
	Surflex	0.814	0.917	0.913	0.906	0.917	0.894	0.904	0.927
	Combined		0.854	0.797	0.853	0.854	0.845	0.882	0.867
FXa	FRED	0.691	0.748	0.743	0.733	0.748	0.747	0.730	0.753
	Surflex	0.583	0.651	0.638	0.673	0.651	0.670	0.614	0.753
	Combined		0.687	0.682	0.665	0.687	0.688	0.668	0.695
TH	FRED	0.586	0.697	0.685	0.613	0.700	0.669	0.700	0.753
	Surflex	0.604	0.771	0.744	0.762	0.771	0.793	0.668	0.848
	Combined		0.614	0.603	0.541	0.614	0.597	0.603	0.756

DUD decoys		Single	Mean	Trim Mean	S1	S2	S3	S4	Sopt
TK	FRED	0.526	0.518	0.516	0.501	0.518	0.510	0.521	0.639
	Surflex	0.514	0.526	0.537	0.500	0.511	0.514	0.538	0.632
	Combined		0.515	0.515	0.498	0.515	0.508	0.518	0.640
GART	FRED	0.652	0.671	0.668	0.536	0.671	0.628	0.702	0.832
	Surflex	0.797	0.879	0.889	0.581	0.715	0.857	0.867	0.892
	Combined		0.850	0.860	0.558	0.689	0.826	0.839	0.861
TH	FRED	0.553	0.534	0.551	0.573	0.534	0.567	0.518	0.639
	Surflex	0.917	0.944	0.944	0.772	0.879	0.926	0.944	0.938
	Combined		0.643	0.631	0.609	0.568	0.636	0.696	0.719
CDK2	FRED	0.601	0.608	0.610	0.586	0.608	0.605	0.607	0.746
	Surflex	0.726	0.766	0.784	0.551	0.606	0.743	0.774	0.707
	Combined		0.613	0.612	0.585	0.606	0.602	0.637	0.766

FXa	FRED	0.565	0.583	0.577	0.491	0.583	0.554	0.596	0.639
	Surflex	0.870	0.907	0.918	0.628	0.750	0.893	0.911	0.918
	Combined		0.587	0.580	0.496	0.589	0.578	0.594	0.636
NA	FRED	0.533	0.544	0.548	0.493	0.544	0.534	0.544	0.764
	Surflex	0.807	0.863	0.863	0.118	0.415	0.709	0.937	0.983
	Combined		0.649	0.666	0.500	0.585	0.614	0.655	0.823
Trypsin	FRED	0.520	0.514	0.513	0.521	0.514	0.513	0.515	0.612
	Surflex	0.988	0.995	0.995	0.913	0.975	0.993	0.988	0.985
	Combined		0.982	0.980	0.891	0.957	0.982	0.962	0.967
COMT	FRED	0.513	0.502	0.492	0.440	0.502	0.481	0.510	0.933
	Surflex	0.517	0.534	0.530	0.475	0.433	0.507	0.550	0.655
	Combined		0.688	0.702	0.570	0.720	0.752	0.618	0.968
ER	FRED	0.678	0.700	0.687	0.602	0.700	0.663	0.716	0.775
	Surflex	0.515	0.507	0.510	0.521	0.510	0.508	0.515	0.625
	Combined		0.837	0.787	0.656	0.820	0.834	0.801	0.847
INHA	FRED	0.515	0.506	0.506	0.468	0.506	0.499	0.512	0.883
	Surflex	0.569	0.591	0.592	0.466	0.486	0.546	0.605	0.648
	Combined		0.567	0.584	0.527	0.685	0.609	0.550	0.904

Table 6. The enrichments of FRED and Surflex for different single target conformations and and combined enrichments) calculated via S_{Mean}, $S_{rimMean}$, S_{opt} at 1 % of the total database.

	*TK*_F	*NA*_F	*FXa*_F	*TH*_F	*TK*_S	*NA*_S	*FXa*_S	*TH*_S
Conf1	13.6	14.3	4.2	1.5	9.1	0	1.4	1.5
Conf2	9	16.3	5.6	1.5	9.1	61.2	1.4	1.5
Conf3	0	2	5.6	1.5	4.5	53.1	4.2	13.8
Conf4	13.6	22.4	8.4	6.2	13.6	0	7	9.2
Conf5	36.4	6.1	3.5	3.1	9.1	59.2	4.2	3.1
Conf6	36.4	8.2	0	3.1	9.1	10.2	0	16.9
Conf7	13.6	12.2	7.7	3.1	9.1	32.7	2.1	1.5
Conf8	9	12.2	1.4	1.5	9.1	0	2.1	3.1
S_{Mean}	27.3	20.4	7.7	7.7	18.2	57.2	3.5	23.1
$S_{TrimMean}$	27.3	20.4	7	9.2	18.2	59.2	3.5	21.5
S_{opt}	31.82	22.45	4.23	12.31	27.27	65.31	7.75	20

TK_F: FRED screened in TK; TK_S: Surflex screened in TK. S_{Mean}, $S_{TrimMean}$, S_{opt}: The enrichment of the docked values of eight conformers by method $S_{Mean,}$ $S_{TrimMean}$ and S_{opt} respectively. Conf1 to 8: The enrichment of the docked value for the individual conformers 1 to 8. S_{Mean}, $S_{TrimMean}$, S_{opt}: The enrichment of the combined docked values of eight conformers by method as calculated by $S_{Mean,}$ $S_{TrimMean}$ and S_{opt} respectively.

Table 7. *p*-Value matrix for 7 consensus methods by calculation of ROC curves and AUC values for 10 targets and decoys from the DUD. 'Single' was calculated as the median of 8 individual ROC curves. Values below 0.1 are in bold to emphasize significance at the 90% level.

		Mean	Trim Mean	S1	S2	S3	S4	Sopt
TK FRED	Single	0.72	0.807	0.868	0.677	0.633	0.863	**0.095**
	Mean		0.837	0.981	0.869	0.808	0.692	**0.084**
	Trim Mean			0.954	0.771	0.725	0.766	**0.089**
	S1				0.948	0.832	0.778	**0.091**
	S2					0.853	0.665	**0.082**
	S3						0.602	**0.079**
	S4							**0.096**
GART FRED	Single	0.194	0.178	0.117	0.195	0.696	0.127	**0.045**
	Mean		0.888	**0.053**	0.965	0.303	0.404	0.123
	Trim Mean			**0.055**	0.901	0.318	0.385	0.116
	S1				**0.053**	0.107	**0.034**	**0.012**
	S2					0.304	0.402	0.122
	S3						0.181	**0.059**
	S4							0.215

TH FRED	Single	0.604	0.836	0.254	0.605	0.401	0.279	**0.03**
	Mean		0.672	0.197	0.978	0.304	0.368	**0.024**
	Trim Mean			0.237	0.674	0.378	0.3	**0.028**
	S1				0.198	0.478	0.112	**0.08**
	S2					0.304	0.367	**0.024**
	S3						0.166	**0.049**
	S4							**0.014**
CDK2 FRED	Single	0.72	0.986	0.219	0.721	0.642	0.702	**0.011**
	Mean		0.677	0.279	0.985	0.867	0.913	**0.009**
	Trim Mean			0.226	0.679	0.66	0.694	**0.01**
	S1				0.278	0.313	0.28	**0.003**
	S2					0.864	0.911	**0.009**
	S3						0.941	**0.008**
	S4							**0.007**
FXa FRED	Single	0.992	0.966	**0.024**	0.996	0.218	0.298	**0.085**
	Mean		0.97	**0.024**	0.972	0.217	0.3	**0.086**
	Trim Mean			**0.023**	0.964	0.213	0.305	**0.087**
	S1				**0.024**	**0.063**	**0.012**	**0.004**
	S2					0.218	0.3	**0.086**
	S3						0.097	**0.032**
	S4							0.185
NA FRED	Single	0.703	0.728	0.265	0.698	0.512	0.825	**0.006**
	Mean		0.909	0.337	0.972	0.685	0.922	**0.005**
	Trim Mean			0.318	0.901	0.634	0.975	**0.006**
	S1				0.338	0.402	0.333	**0.003**
	S2					0.688	0.918	**0.005**
	S3						0.694	**0.005**
	S4							**0.006**
Trypsin FRED	Single	0.899	0.898	0.508	0.908	0.824	0.836	**0.007**
	Mean		0.987	0.538	0.949	0.901	0.903	**0.007**
	Trim Mean			0.538	1	0.891	0.899	**0.007**
	S1				0.533	0.564	0.569	**0.01**
	S2					0.893	0.895	**0.007**
	S3						0.997	**0.007**
	S4							**0.007**
COMT FRED	Single	0.742	0.834	0.43	0.766	0.663	0.445	**0.011**
	Mean		0.819	0.39	0.926	0.565	0.507	**0.012**
	Trim Mean			0.408	0.851	0.607	0.471	**0.011**
	S1				0.396	0.534	0.267	**0.007**
	S2					0.574	0.502	**0.012**
	S3						0.374	**0.01**
	S4							**0.016**
ER FRED	Single	0.822	0.98	**0.09**	0.828	0.282	0.499	0.17
	Mean		0.786	0.1	0.966	0.318	0.427	0.158
	Trim Mean			0.092	0.792	0.284	0.489	0.172
	S1				0.1	0.204	0.058	**0.023**
	S2					0.317	0.43	0.159
	S3						0.182	**0.074**
	S4							0.263
	Single	0.884	0.964	0.194	0.883	0.548	0.565	**0**
	Mean		0.873	0.209	0.994	0.58	0.514	**0**

INHA FRED	Trim Mean			0.199	0.872	0.548	0.549	**0**
	S1				0.209	0.288	0.136	**0**
	S2					0.58	0.514	**0**
	S3						0.363	**0**
	S4							**0**
TK Surflex	Single	0.500	0.542	0.378	0.461	0.417	0.865	**0.082**
	Mean		0.803	0.583	0.715	0.702	0.616	0.120
	Trim Mean			0.528	0.648	0.626	0.721	0.112
	S1				0.732	0.752	0.418	0.163
	S2					0.943	0.502	0.141
	S3						0.498	0.141
	S4							**0.094**
GART Surflex	Single	**0.058**	**0.012**	**0.061**	0.270	0.421	0.206	**0.091**
	Mean		0.806	**0.019**	**0.082**	0.566	0.786	0.761
	Trim Mean			**0.018**	**0.077**	0.501	0.681	0.857
	S1				0.122	**0.025**	**0.021**	**0.017**
	S2					0.108	**0.091**	**0.071**
	S3						0.738	0.455
	S4							0.584
TH Surflex	Single	**0.030**	**0.011**	**0.023**	0.275	0.723	0.212	0.209
	Mean		0.934	**0.007**	**0.078**	0.485	0.778	0.960
	Trim Mean			**0.007**	**0.078**	0.491	0.831	0.930
	S1				**0.041**	**0.01**	0.008	**0.003**
	S2					0.121	0.093	**0.034**
	S3						0.585	0.391
	S4							0.822
CDK2 Surflex	Single	0.285	**0.061**	**0.038**	**0.098**	0.814	0.222	0.559
	Mean		0.535	0	**0.002**	0.375	0.717	**0.065**
	Trim Mean			0	**0.002**	0.259	0.887	**0.042**
	S1				0.069	**0.001**	**0**	**0.003**
	S2					**0.005**	**0.002**	**0.021**
	S3						0.34	0.212
	S4							**0.046**
FXa Surflex	Single	**0.058**	**0.048**	0.329	0.826	**0.082**	**0.053**	**0.042**
	Mean		0.937	**0.004**	**0.095**	0.879	0.966	0.887
	Trim Mean			**0.003**	**0.080**	0.817	0.970	0.950
	S1				0.232	**0.007**	**0.004**	**0.003**
	S2					0.129	**0.087**	**0.070**
	S3						0.846	0.769
	S4							0.921
NA Surflex	Single	0.172	0.178	**0**	**0.002**	0.807	**0.029**	**0.007**
	Mean		0.892	**0**	**0**	**0.02**	0.132	**0.047**
	Trim Mean			**0**	**0**	**0.021**	0.126	**0.046**
	S1				**0.002**	**0**	**0**	**0**
	S2					**0.002**	**0**	**0**
	S3						**0.005**	**0.002**
	S4							0.215
Trypsin Surflex	Single	0.605	0.617	0.667	0.799	0.609	0.747	0.709
	Mean		0.778	**0.023**	0.256	0.956	0.273	0.399
	Trim Mean			**0.025**	0.28	0.92	0.293	0.444

	S1				0.052	0.022	0.052	0.036
	S2					0.24	0.723	0.46
	S3						0.297	0.372
	S4							0.733
COMT Surflex	Single	0.801	0.797	0.602	0.322	0.635	0.883	**0.086**
	Mean		0.978	0.730	0.404	0.608	0.814	0.189
	Trim Mean			0.732	0.403	0.601	0.799	0.189
	S1				0.443	0.950	0.603	0.192
	S2					0.530	0.324	0.331
	S3						0.540	0.240
	S4							0.159
ER Surflex	Single	0.900	0.969	0.629	0.991	0.981	0.844	**0.094**
	Mean		0.855	0.563	0.862	0.924	0.776	0.101
	Trim Mean			0.635	0.980	0.955	0.868	**0.095**
	S1				0.612	0.605	0.738	**0.075**
	S2					0.971	0.857	**0.094**
	S3						0.855	**0.099**
	S4							**0.089**
INHA Surflex	Single	0.752	0.747	0.331	0.447	0.882	0.627	0.383
	Mean		0.960	**0.020**	**0.039**	0.198	0.472	0.158
	Trim Mean			**0.020**	**0.039**	0.198	0.485	0.163
	S1				0.399	**0.060**	**0.014**	**0.005**
	S2					0.115	**0.028**	**0.008**
	S3						0.134	**0.04**
	S4							0.312
TK Combined	Mean		0.683	0.299	0.44	0.507	0.463	**0.018**
	Trim Mean			0.255	0.365	0.423	0.573	**0.016**
	S1				0.526	0.465	0.196	**0.034**
	S2					0.850	0.269	**0.026**
	S3						0.312	**0.024**
	S4							**0.013**
GART Combined	Mean		0.671	**0.001**	**0.014**	0.364	0.606	0.518
	Trim Mean			**0.001**	**0.012**	0.296	0.470	0.683
	S1				**0.025**	**0.001**	**0.001**	**0.001**
	S2					**0.023**	**0.017**	**0.010**
	S3						0.560	0.246
	S4							0.343
TH Combined	Mean		0.831	**0**	**0.011**	0.193	0.658	0.181
	Trim Mean			**0**	**0.012**	0.205	0.711	0.191
	S1				**0.003**	**0**	**0**	**0**
	S2					**0.031**	**0.012**	**0.032**
	S3						0.227	0.818
	S4							0.207
CDK2 Combined	Mean		0.28	**0**	**0**	0.112	0.424	**0.007**
	Trim Mean			**0**	**0**	**0.052**	0.745	**0.003**
	S1				**0.008**	**0**	**0**	**0**
	S2					**0**	**0**	**0.001**
	S3						**0.067**	**0.037**
	S4							**0.004**
	Mean		0.339	0	0	0.182	0.497	0.191
	Trim			0	0	0.099	0.587	0.391

FXa Combined	Mean							
	S1				**0**	**0**	**0**	**0**
	S2					**0**	**0**	**0**
	S3						0.125	**0.060**
	S4							0.283
NA Combined	Mean		0.782	**0**	**0**	**0.001**	**0.020**	**0.003**
	Trim Mean			**0**	**0**	**0.001**	**0.019**	**0.003**
	S1				**0**	**0**	**0**	**0**
	S2					**0**	**0**	**0**
	S3						**0**	**0**
	S4							**0.062**
Trypsin Combined	Mean		0.741	**0.001**	**0.080**	0.914	0.189	0.211
	Trim Mean			**0.001**	**0.086**	0.867	0.205	0.229
	S1				**0.003**	**0.001**	**0.002**	**0.001**
	S2					**0.090**	0.436	0.227
	S3						0.213	0.240
	S4							0.617
COMT Combined	Mean		0.939	0.399	0.119	0.398	0.632	**0.030**
	TrimMean			0.409	0.122	0.408	0.601	**0.030**
	S1				0.229	0.867	0.276	**0.053**
	S2					0.203	**0.094**	0.135
	S3						0.309	**0.047**
	S4							**0.024**
ER Combined	Mean		0.730	0.355	0.757	0.835	0.575	**0.017**
	TrimMean			0.417	0.962	0.904	0.717	**0.015**
	S1				0.399	0.388	0.526	**0.010**
	S2					0.932	0.695	**0.015**
	S3						0.681	**0.016**
	S4							**0.013**
INHA Combined	Mean		0.910	**0.001**	**0.002**	**0.047**	0.238	**0.027**
	TrimMean			**0.001**	**0.002**	**0.046**	0.246	**0.028**
	S1				0.184	**0.008**	**0**	**0**
	S2					**0.022**	**0.001**	**0**
	S3						**0.021**	**0.003**
	S4							**0.071**

Acknowledgements

We thank the Bayer Haemophilia Awards programme for their support with a special project grant to G.A.F.N. and furthermore the China Scholarship Council (grant no. 2008630114).

References

1. Mandal S, Moudgil M, & Mandal SK (2009) Rational drug design. *Eur J Pharmacol* 625(1-3):90-100.
2. Mavromoustakos T, *et al.* (2011) Strategies in the rational drug design. *Curr Med Chem* 18(17):2517-2530.
3. Irwin JJ & Shoichet BK (2005) ZINC--a free database of commercially available compounds for virtual screening. *J Chem Inf Model* 45(1):177-182.
4. Clark DE (2006) What has virtual screening ever done for drug discovery? *Expert Opinion on Drug Discovery* 1:103-110.
5. Reddy AS, Pati SP, Kumar PP, Pradeep HN, & Sastry GN (2007) Virtual screening in drug discovery -- a computational perspective. *Current protein & peptide science* 8(4):329-351.
6. Lengauer T, Lemmen C, Rarey M, & Zimmermann M (2004) Novel technologies for virtual screening. *Drug discovery today* 9(1):27-34.
7. Quintus F, Sperandio O, Grynberg J, Petitjean M, & Tuffery P (2009) Ligand scaffold hopping combining 3D maximal substructure search and molecular similarity. *BMC bioinformatics* 10:245.
8. Mason JS, Good AC, & Martin EJ (2001) 3-D pharmacophores in drug discovery. *Current pharmaceutical design* 7(7):567-597.
9. Srinivasan J, *et al.* (2002) Evaluation of a novel shape-based computational filter for lead evolution: application to thrombin inhibitors. *Journal of medicinal chemistry* 45(12):2494-2500.
10. Muchmore SW, Souers AJ, & Akritopoulou-Zanze I (2006) The use of three-dimensional shape and electrostatic similarity searching in the identification of a melanin-concentrating hormone receptor 1 antagonist. *Chemical biology & drug design* 67(2):174-176.
11. Rush TS, 3rd, Grant JA, Mosyak L, & Nicholls A (2005) A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. *Journal of medicinal chemistry* 48(5):1489-1495.
12. Ballester PJ, Finn PW, & Richards WG (2009) Ultrafast shape recognition: evaluating a new ligand-based virtual screening technology. *Journal of molecular graphics & modelling* 27(7):836-845.
13. Sastry GM, Dixon SL, & Sherman W (2011) Rapid shape-based ligand alignment and virtual screening method based on atom/feature-pair similarities and volume overlap scoring. *Journal of chemical information and modeling* 51(10):2455-2466.
14. Yuriev E, Agostino M, & Ramsland PA (2011) Challenges and advances in computational docking: 2009 in review. *Journal of molecular recognition : JMR* 24(2):149-164.

15. Villoutreix BO, Eudes R, & Miteva MA (2009) Structure-based virtual ligand screening: recent success stories. *Combinatorial chemistry & high throughput screening* 12(10):1000-1016.
16. Perola E & Charifson PS (2004) Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding. *Journal of medicinal chemistry* 47(10):2499-2510.
17. Giganti D, *et al.* (2010) Comparative evaluation of 3D virtual ligand screening methods: impact of the molecular alignment on enrichment. *Journal of chemical information and modeling* 50(6):992-1004.
18. McNally VA, *et al.* (2003) Identification of a novel class of inhibitor of human and Escherichia coli thymidine phosphorylase by in silico screening. *Bioorganic & medicinal chemistry letters* 13(21):3705-3709.
19. Mysinger MM, *et al.* (2012) Structure-based ligand discovery for the protein-protein interface of chemokine receptor CXCR4. *Proceedings of the National Academy of Sciences of the United States of America* 109(14):5517-5522.
20. Gehlhaar DK, *et al.* (1995) Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming. *Chemistry & biology* 2(5):317-324.
21. Dey R & Chen L (2011) In search of allosteric modulators of a7-nAChR by solvent density guided virtual screening. *Journal of biomolecular structure & dynamics* 28(5):695-715.
22. McGann MR, Almond HR, Nicholls A, Grant JA, & Brown FK (2003) Gaussian docking functions. *Biopolymers* 68(1):76-90.
23. McGann M (2011) FRED pose prediction and virtual screening accuracy. *Journal of chemical information and modeling* 51(3):578-596.
24. Jain AN (2003) Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *Journal of medicinal chemistry* 46(4):499-511.
25. Friesner RA, *et al.* (2004) Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *Journal of medicinal chemistry* 47(7):1739-1749.
26. Halgren TA, *et al.* (2004) Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *Journal of medicinal chemistry* 47(7):1750-1759.
27. Liebeschuetz JW, Cole JC, & Korb O (2012) Pose prediction and virtual screening performance of GOLD scoring functions in a standardized test. *J Comput Aided Mol Des*.
28. Neves MA, Totrov M, & Abagyan R (2012) Docking and scoring with ICM: the benchmarking results and strategies for improvement. *J Comput Aided Mol Des*.
29. Venkatachalam CM, Jiang X, Oldfield T, & Waldman M (2003) LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. *Journal of molecular graphics & modelling* 21(4):289-307.
30. Villoutreix BO, *et al.* (2007) Free resources to assist structure-based virtual ligand screening experiments. *Current protein & peptide science* 8(4):381-411.
31. Yuriev E & Ramsland PA (2013) Latest developments in molecular docking: 2010-2011 in review. *Journal of molecular recognition : JMR* 26(5):215-239.
32. Miteva MA, Lee WH, Montes MO, & Villoutreix BO (2005) Fast structure-based virtual ligand screening combining FRED, DOCK, and Surflex. *Journal of medicinal chemistry* 48(19):6012-6022.

33. Cozza G, *et al.* (2006) Identification of ellagic acid as potent inhibitor of protein kinase CK2: a successful example of a virtual screening application. *Journal of medicinal chemistry* 49(8):2363-2366.
34. Villoutreix BO, *et al.* (2011) Tyrosine kinase syk non-enzymatic inhibitors and potential anti-allergic drug-like compounds discovered by virtual and in vitro screening. *PloS one* 6(6):e21117.
35. Teague SJ (2003) Implications of protein flexibility for drug discovery. *Nat Rev Drug Discov* 2(7):527-541.
36. Bottegoni G, Rocchia W, Rueda M, Abagyan R, & Cavalli A (2011) Systematic exploitation of multiple receptor conformations for virtual ligand screening. *PLoS One* 6(5):e18845.
37. C BR, Subramanian J, & Sharma SD (2009) Managing protein flexibility in docking and its applications. *Drug Discov Today* 14(7-8):394-400.
38. Cozzini P, *et al.* (2008) Target flexibility: an emerging consideration in drug discovery and design. *J Med Chem* 51(20):6237-6255.
39. Totrov M & Abagyan R (2008) Flexible ligand docking to multiple receptor conformations: a practical alternative. *Curr Opin Struct Biol* 18(2):178-184.
40. Cavasotto CNO, A.J.W. and Abagyan, R. (2005) The Challenge of Considering Rreceptor Flexibility in Ligand Docking and Virtual Screening. *Curr. Comput.-Aided Drug Des* 1:423-440
41. Craig IR, Essex JW, & Spiegel K (2010) Ensemble docking into multiple crystallographically derived protein structures: an evaluation based on the statistical analysis of enrichments. *Journal of chemical information and modeling* 50(4):511-524.
42. Osguthorpe DJ, Sherman W, & Hagler AT (2012) Exploring protein flexibility: incorporating structural ensembles from crystal structures and simulation into virtual screening protocols. *The journal of physical chemistry. B* 116(23):6952-6959.
43. Barbara Sander, Oliver Korb, Jason Cole, & Essex JW (2012) The assessment of computationally derived protein ensembles in protein-ligand docking. *Journal of cheminformatics*.
44. Cavasotto CN (2012) Normal mode-based approaches in receptor ensemble docking. *Methods in molecular biology* 819:157-168.
45. Charifson PS, Corkery JJ, Murcko MA, & Walters WP (1999) Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *Journal of medicinal chemistry* 42(25):5100-5109.
46. Plewczynski D, Lazniewski M, von Grotthuss M, Rychlewski L, & Ginalski K (2011) VoteDock: consensus docking method for prediction of protein-ligand interactions. *Journal of computational chemistry* 32(4):568-581.
47. Sperandio O, *et al.* (2010) How to choose relevant multiple receptor conformations for virtual screening: a test case of Cdk2 and normal mode analysis. *Eur Biophys J* 39(9):1365-1372.
48. Rueda M, Bottegoni G, & Abagyan R (2010) Recipes for the selection of experimental protein conformations for virtual screening. *J Chem Inf Model* 50(1):186-193.
49. Suhre K, Navaza J, & Sanejouand YH (2006) NORMA: a tool for flexible fitting of high-resolution protein structures into low-resolution electron-microscopy-derived density maps. *Acta Crystallogr D Biol Crystallogr* 62(Pt 9):1098-1100.

50. Huang N, Kalyanaraman C, Irwin JJ, & Jacobson MP (2006) Physics-based scoring of protein-ligand complexes: enrichment of known inhibitors in large-scale virtual screening. *J Chem Inf Model* 46(1):243-253.
51. Abagyan R & Totrov M (1994) Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. *J Mol Biol* 235(3):983-1002.
52. Levitt M, Sander C, & Stern PS (1985) Protein normal-mode dynamics: trypsin inhibitor, crambin, ribonuclease and lysozyme. *J Mol Biol* 181(3):423-447.
53. Suhre K & Sanejouand YH (2004) ElNemo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement. *Nucleic acids research* 32(Web Server issue):W610-614.
54. Huang N, Shoichet BK, & Irwin JJ (2006) Benchmarking sets for molecular docking. *Journal of medicinal chemistry* 49(23):6789-6801.
55. Chuprina A, Lukin O, Demoiseaux R, Buzko A, & Shivanyuk A (2010) Drug- and lead-likeness, target class, and molecular diversity analysis of 7.9 million commercially available organic compounds provided by 29 suppliers. *J Chem Inf Model* 50(4):470-479.
56. Hawkins PC, Skillman AG, Warren GL, Ellingson BA, & Stahl MT (2010) Conformer generation with OMEGA: algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database. *Journal of chemical information and modeling* 50(4):572-584.
57. Hawkins PC & Nicholls A (2012) Conformer generation with OMEGA: learning from the data set and the analysis of failures. *Journal of chemical information and modeling* 52(11):2919-2936.
58. Joosten RP, Joosten K, Murshudov GN, & Perrakis A (2012) PDB_REDO: constructive validation, more than just looking for errors. *Acta Crystallogr D Biol Crystallogr* 68(Pt 4):484-496.
59. Eldridge MD, Murray CW, Auton TR, Paolini GV, & Mee RP (1997) Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J Comput Aided Mol Des* 11(5):425-445.
60. DeLano WL (2002) The PyMOL Molecular Graphics System.
61. Welch W, Ruppert J, & Jain AN (1996) Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites. *Chem Biol* 3(6):449-462.
62. Ihaka R & Gentleman R (1996) R: A Language for Data Analysis and Graphics. *J Comp Graph Stat* 5(3):299-314.
63. Bender A & Glen RC (2005) A discussion of measures of enrichment in virtual screening: comparing the information content of descriptors with increasing levels of sophistication. *J Chem Inf Model* 45(5):1369-1375.
64. Oprea TI, Davis AM, Teague SJ, & Leeson PD (2001) Is there a difference between leads and drugs? A historical perspective. *Journal of chemical information and computer sciences* 41(5):1308-1315.
65. Hammes GG (2002) Multiple conformational changes in enzyme catalysis. *Biochemistry* 41(26):8221-8228.
66. Nussinov R & Ma B (2012) Protein dynamics and conformational selection in bidirectional signal transduction. *BMC Biol* 10:2.
67. Oprea TI, Davis AM, Teague SJ, & Leeson PD (2001) Is there a difference between leads and drugs? A historical perspective. *Journal of chemical information and computer sciences* 41(5):1308-1315.

Chapter 3

Molecular simulation studies of human coagulation factor VIII C domain-mediated membrane binding

Jiangfeng Du, Kanin Wichapong, Tilman M. Hackeng,

Gerry A.F. Nicolaes

Submitted to Thrombosis and Haemostasis

Abstract

The C-terminal C domains of activated coagulation factor VIII (FVIIIa) are essential to membrane binding of this crucial coagulation cofactor protein. To provide an overall membrane binding mechanism for FVIII, we have performed simulations of membrane binding through coarse-grained molecular dynamics simulations of the C1 and C2 domain, and the combined C-domains (C1+C2). We have found that the C1 and C2 domain have different membrane binding properties. The C1 domain uses hydrophobic spikes 3 and 4, of its total of four spikes, as major loops to bind the membrane, whereas all four of its hydrophobic loops of the C2 domain appear essential for membrane binding. Interestingly, in the C1+C2 system, we observe cooperative binding of the C1 and C2 domains such that all four C2 domain spikes bound first, after which all four loops of the C1 domain inserted into the membrane, while the net binding energy is higher than that of the sum of the isolated C domains. Several residues, mutation of which are known to cause Haemophilia A, were identified as key residues for membrane binding. In addition to these known residues, we identified residues from the C1 and C2 domains which are involved in the membrane binding process, that have not been reported before as a cause for haemophilia A, but which contribute to overall membrane binding and which are likely candidates for novel causative missense mutations in haemophilia A.

Keywords Coarse-Grained Molecular Dynamics (CGMD), Coagulation FVIII, C domains, membrane, haemophilia A

Introduction

Blood coagulation factor VIII (FVIII) is the procofactor of FVIIIa, the non-enzymatic cofactor of the intrinsic factor Xase complex. FVIII has a mosaic structure and consists of 6 domains: A1-A2-B-A3-C1-C2. The A1, A2 and A3 domains share 40% sequence similarity while the C1 and C2 domain are 54% similar (1). The carboxyl terminal C domains are involved in binding to von Willebrand factor (VWF) and to negatively charged membrane surfaces (2-6). The C domains are homologous to each other and to the coagulation factor V (FV) C domains (7).

In blood, FVIII circulates in complex with VWF and thrombin-catalyzed activation of FVIII is associated with limited proteolysis of FVIII at R372, R740, and R1689. After cleavage at R1689, VWF dissociates from FVIII whereas the B chain is released after cleavage at R740 (8). FVIIIa forms the intrinsic tenase complex with activated coagulation factor IX (FIXa) on a phospholipid surface and converts coagulation factor X (FX) to its activated form (FXa). Deficiency of FVIII is defined as haemophilia A, a sex-linked disease. FVIIIa binding to negatively charged membrane surfaces enhances the conversion of FX to FXa by over 200,000-fold (9-11).

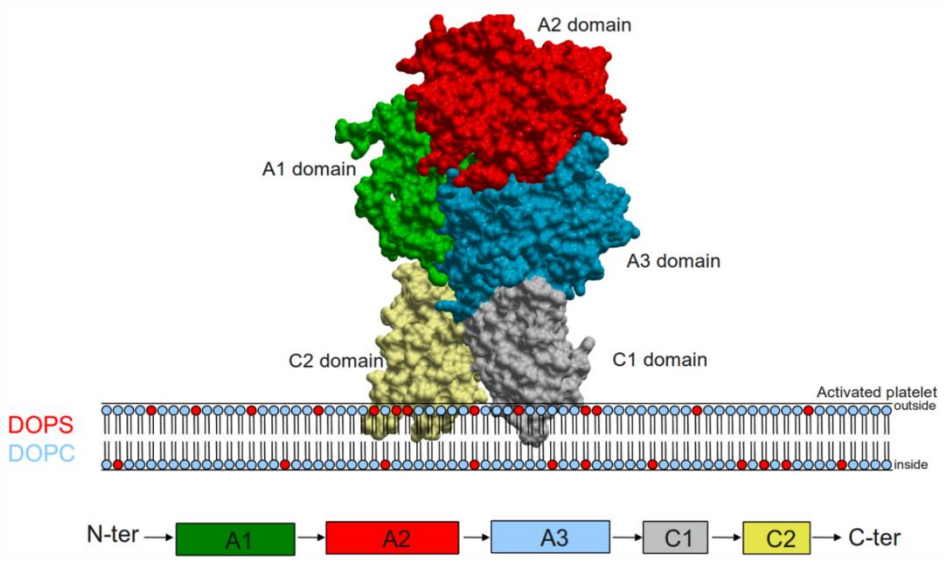


Figure 1. Model of membrane-bound FVIIIa. FVIIIa contains five different domains that are organized from N-to C terminus: A1-A2-A3-C1-C2. The 3D representation of FVIIIa was based on the crystallographic structure of human recombinant FVIII deposited as pdb entry file 2R7E.pdb. FVIIIa is composed of a heavy chain (the A1 (Green) and A2 (Red) domains) and light chain (A3 (Cyan), C1 (Gray) and C2 (Yellow) domains). FVIIIa binds to activated membranes via its C1 and C2 domains with the FVIIIa A domains here positioned on the top of the C domains.

The FVIIIa C1 domain (residue C2021-C2169) and C2 domain (residue L2171-Q2329) are the main membrane-binding regions of the cofactor (Figure 1), of which the C2 domain has been hypothesized as being most important (2-6, 12, 13). The homologous FVIII C1 and C2 domains are highly similar in 3D structure (Figure 2): both of them are typical beta barrels, where anti-parallel sheets are linked by four hydrophobic spikes (shown and defined in Figure 2), which are hypothesized to bind to the membrane (13-15). Both C-domains are basic, their pI being 9.96 and 9.43 respectively and it is believed that the membrane approach is facilitated by electrostatic interactions between the positively charged C-domains and the net-

negatively charged membrane (14).

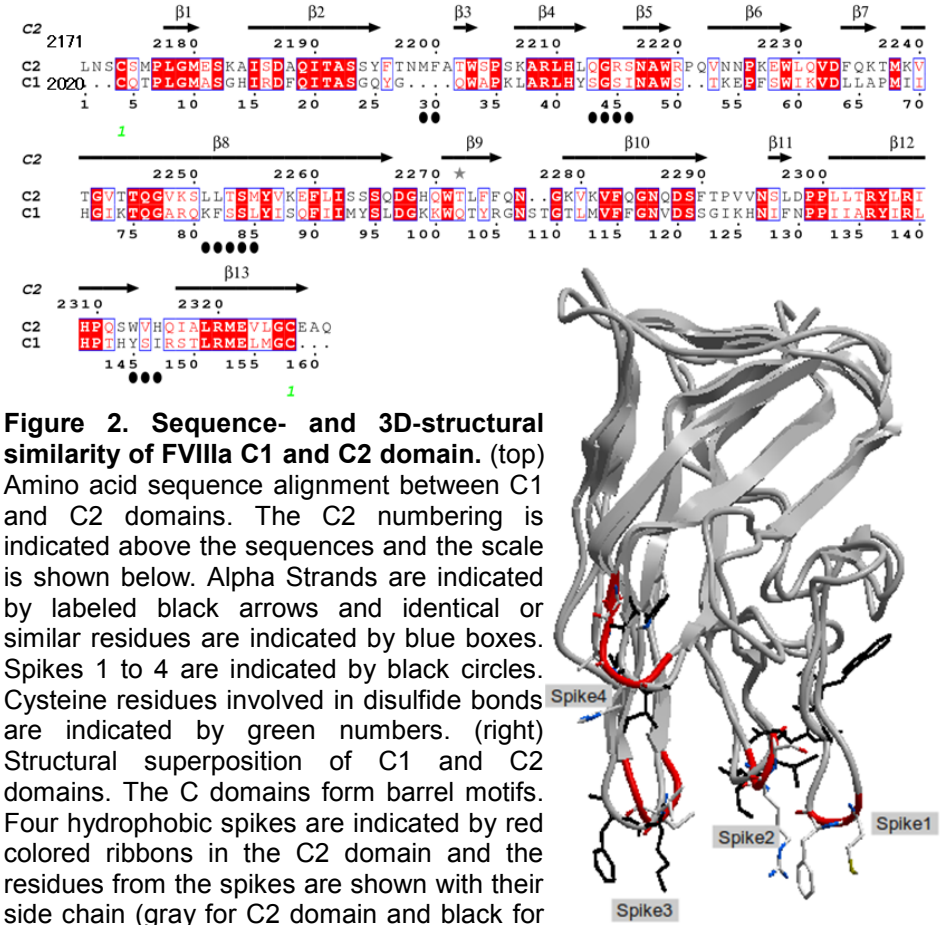


Figure 2. Sequence- and 3D-structural similarity of FVIIIa C1 and C2 domain. (top) Amino acid sequence alignment between C1 and C2 domains. The C2 numbering is indicated above the sequences and the scale is shown below. Alpha Strands are indicated by labeled black arrows and identical or similar residues are indicated by blue boxes. Spikes 1 to 4 are indicated by black circles. Cysteine residues involved in disulfide bonds are indicated by green numbers. (right) Structural superposition of C1 and C2 domains. The C domains form barrel motifs. Four hydrophobic spikes are indicated by red colored ribbons in the C2 domain and the residues from the spikes are shown with their side chain (gray for C2 domain and black for Four spikes for the FVIII C1 and C2 domains can be defined: spike 1, G2044-Q2045; spike2, Y2055-I2059; spike3, K2092-F2093 and spike4, I2158-R2159 in the FVIII C1 domain and spike 1, M2199-F2200; spike2, Q2213-S2216; spike3, L2251-M2255 and spike4, W2313-H2315 for the C2 domain.

The exact binding mechanism of the FVIII C domains remains unknown. Several studies illustrated the likelihood of the membrane interaction interface being formed by the C1 and the C2 domain (16-18), while others indicated that the C2 domain is the sole membrane-binding motif (19). Crystal structures have been solved that provide detailed information on the 3D structure of the C1 domain, the C2

domain and on the structure of a complete human factor VIIIa molecule. The latter structure suggests that the C1 and C2 domains are juxtaposed on a membrane surface (16-18). Cryo-electron microscopy (EM) presented an alternative model of FVIIIa, in which only the C2 domain is bound to the membrane while the C1 domain is superpositioned to the C2 domain (19). Autin and co-workers employed homology modeling to build ten representative models of the FIXa-FVIIIa complex. In their most representative model 5, the C1 and A3 domains positioned in one plane parallel to the membrane surface and the C2 domain binding to the membrane (20). Later, Ngo and coworkers proposed a homology model of the FIXa-FVIIIa complex with both C1 and C2 domains and a loop (C1899-C1903) from the A3 domain to interact with the membrane (16). Stoilova-McPhie and co-workers applied 2D electron microscopy to study the membrane-bound FVIIIa structure and concluded that the A3, C1 and C2 domains were near the membrane surface, with the C2 slightly inclined to the surface while the C1 orientation was constrained during the EM experiments (21). Study of the C domain structures of FVIII (7, 16), has resulted in the hypothesis that hydrophobic spikes (as shown in Figure 2) play an important role in membrane binding. The importance of the hydrophobic spikes has been experimentally confirmed by mutagenesis studies (13, 22). Several causative mutations for haemophilia A have been described to reside in the FVIII C-domains. In fact, 13.9 %, of all hemophilia A -associated missense mutations, as collected in the haemophilia A database (HAMSTeRS, ref (23) originate from the C domains. These mutations link the membrane binding properties of FVIII to haemophilia A. A

number of studies have experimentally addressed causal relationships between C-domain missense mutations and haemophilia A (24, 25), these studies can be hampered however by failure to express sufficient recombinant FVIII, as in the case of a W2313R mutation (19, 26, 27). However, a general mechanism that explains all hemophilia-associated C-domain missense mutations and is able to predict functional ramifications of novel mutations is currently not available.

To advance our understanding of the mechanism of FVIIIa C domain-mediated membrane binding, we have applied coarse-grained molecular dynamic simulations (CGMD) (28) of the FVIII C domains, as a model for the overall FVIIIa membrane binding process. With this study we aim to study the membrane-binding mechanism of the C domains and to identify key residues from C1 and C2 domain which play important roles in membrane binding.

We have analysed the binding mechanism of solely C1 and C2 domains and also C1 together with C2 domain, in order to investigate a potential cooperativity between these two domains in the membrane binding process. To our knowledge, this is the first study focusing on the membrane binding mechanism of both C domains of FVIII by means of CGMD. The results derived from our work provide a detailed molecular framework for the membrane binding process of FVIII. We have identified key residues for membrane binding, some of which have been described in haemophilia A patients, and others which we propose to be candidates for novel causative mutations in haemophilia A carriers.

Materials and Methods

Coarse grained model construction

The 3D structures of the FVIII C1 and C2 domains were retrieved from the Protein Data Bank (PDB Code 3CDZ) and the YASARA/Whatif twin package (29) was used to add hydrogen atoms according to the protonation state at pH = 7 and to generate C1/C2 mutants (M2199W/F2200W (C2-spike1 mutant A), M2199W/F2200W/L2252S (C2-spike1/3 mutant A), L2252S (C2-spike 3 mutant), W2313R (C2-spike4 mutant), M2199A/F2200A/L2251A/L2252A (C2-spike 1/3 mutant B) and K2092A/F2093A (C1-spike 3 mutant) in which selected amino acids were swapped. To create the A2201 deletion mutant, we deleted amino acid A2201, and built a homology model based on the FVIII C2 3D structure (retrieved from 3CDZ). All models were optimized by a 3 ns MD simulation with the yasara2 force field in water (29). The coarse-grained models of C domains and their mutants were generated by the martinize script version 2.3 (http://md.chem.rug.nl/cgmartini/index.php/downloads/tools/204-martinize) combined with DSSP2.0.4 (30, 31) and elastic networks (32).

Coarse-grained molecular dynamics (CGMD) simulation

The C domain coarse-grained models were combined with phospholipid membranes of variable composition, in solution at an initial distance of 30 Å to the membrane. To investigate the influence of the DOPS/DOPC membrane ratio on FVIIIa C2 membrane

binding, we have simulated 8 systems in which the FVIIIa C2 domain was allowed to interact with DOPC membranes containing a varying percentage of DOPS of 0%, 2%, 5%, 10%, 15%, 20%, 30% or 50%.To gain insight into the binding mechanism of C1, C2 and the potential cooperativity of C1 and C2 for binding with membrane, we have set up simulations for the isolated C1 or C2 as well as C1 together with C2 domain (C1+C2) under the same conditions (10/90 DOPS/DOPC (mol/mol) membrane). The correctness of our simulation approach was validated by inclusion of several FVIII-C2 variants that are known to bind differently to phospholipid membranes. To this end we selected mutants from several of the spikes from the C1 and C2 domain involved in membrane binding as summarized in Table 1. We selected FVIII variants that possess both improved or reduced membrane binding properties. Also, we included W2313R (C2-spike 4 mutant), a molecule which has been implicated in haemophilia A, but the functional characterization of which could not be addressed experimentally (27).

CGMD simulations of the abovementioned simulation systems (18 in total) were carried out as follows. The Martini v2.2 force field (33) implemented in the Gromacs-4.5.3 package was applied for proteins and membrane. The area of each membrane was set to 0.65 nm^2 per lipid and the height of the simulation box was 180 Å. Counter ions (Na^+ or Cl^-) were added to neutralize the systems. Prior to the CGMD simulations, energy minimization was performed to remove bad contacts which may cause unstable simulations. Next, 50,000 steps of steepest descent with position constraint (1000 kJ/mol) assigned to all protein atoms were applied in this step. Subsequently,

equilibrated MD simulations were performed for 200 ns with 20 femtoseconds (fs) time steps. The temperature was kept constant at 325K by application of Berendsen coupling methods. A semi-isotropic coupling with a compressibility of $1e^{-5}$ bar^{-1} was used to maintain the pressure of the system at 1 bar. A position constraint of 1000 kJ/mol was assigned to all protein atoms and motion of center of mass of membrane was removed. Finally, free CGMD simulations were performed for 2 μs using the NPT ensemble. For some systems, e.g. the C1 system in which this domain did not bind the membrane during up to 2 μs, the simulation was prolonged till 4 μs. A time step was set at 20 fs and the temperature was maintained at 325 K by using Berendsen coupling methods as in the equilibrated MD phase. The pressure in all the simulations was coupled by a semi-isotropic Berendsen method, with a compressibility of $3e^{-4}$. Coordinates and energy of the systems were collected every 5000 steps of simulations. The electrostatics interaction and van der Waals (vdW) interaction were calculated with a shift cutoff (33-35) approach and a Lennard-Jones potential was used for vdW calculation. The multiple body cut-off distance was set to 1.5 nm (32, 33, 35). In short, after an initial equilibration step (0.2 μs), we performed a 2-8 μs simulation of the different C domains in solution, in the presence of a membrane surface of chosen composition, while controlling the pH, temperature and ionic strength of the simulation cell. The MD simulations allowed us to calculate the energy of binding between C domains and the membrane. The binding free energy (ΔG_{bind}) of each system was derived from the potential of mean force (PMF) which was calculated from a series of umbrella sampling simulations. An umbrella sampling

simulation as described in literature (35-37) was performed by a 10 ns steered molecular dynamics simulation, where a constant force (7000kJ/mol/nm) was applied to harmonically pull C domains away from the membrane. From the reaction coordinate, approximately 35 conformations of the C domain and membrane complexes were extracted for umbrella sampling simulation. Those 35 samples located at a center of mass (COM) distance of 6 Å along the reaction coordinate. The 35 samples were further simulated by 10 ns conventional molecular dynamics simulations. The potential mean force (PMF) for each system was composed from 35 samples by means of WHAM algorithm (35, 38). With a constant pulling force and 35 samples, the sample windows overlapped such that the PMF curve could be constructed. The derived PMF curve represents the free energy profile along the reaction coordinate, thus in this case the binding free energy (ΔGbind) can be approximately calculated as the difference of energy between minimum point and the plateau region of the PMF curve. We verified that the initial orientation of the C domain towards the membrane surface did not influence the binding mechanism and that the binding process itself was reproducible. All simulations were repeated at least 3 times.

Analysis of the membrane binding process

During the simulation, the location of every residue was monitored. The distance of a C domain to the membrane is defined as the minimum distance of any coarse-grained ‘atom’ in the domain to any of the membrane atoms. We defined binding events to occur when the minimum distance between a C domain and the membrane was

less than 5 Å and the corresponding simulation time as the membrane-binding time. The orientation of C1, C2 and C1+C2 was monitored by measurement of tilting angles as defined in Figure 3. The tilting angle of each system (C1, C2 or C1+C2) was calculated by averaging the angles from the last 50 snapshots of the simulations. To calculate the free energy between individual residues during membrane binding and the membrane, each residue term was put in the option “energygrps” in the parameter file. The energy state was saved every 100 ps. The free energy calculations were divided into two groups. One is in the approaching stage, where the importance of individual residues to drive the C domains to the membrane can be detected. The second group is in the anchored stage, where the importance of individual residues for interaction with the membrane can be detected.

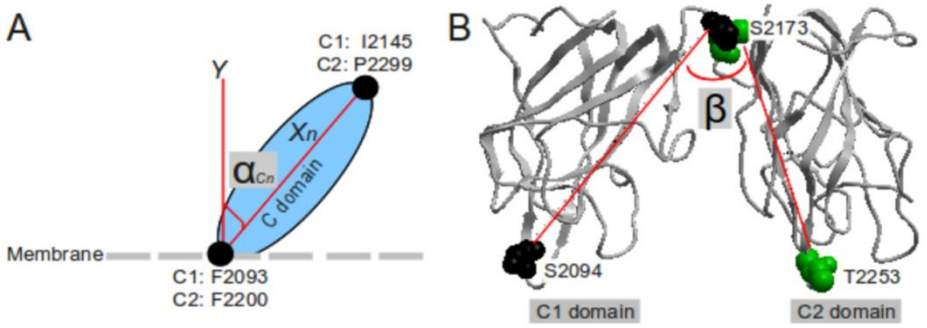


Figure 3. Angle definitions. A. alpha (α) angle is defined as the angle between the line which is perpendicular to the membrane surface and the line through the alpha carbons of F2093 (F2200) and I2145 (P2299) in C1 (C2) domain. B. beta (β) angle is defined as the angle between four alpha carbons of residues (S2094, S2173, C2174 and T2253).

Results

Influence of initial C domain orientation on membrane binding and reproducibility of binding

To test whether the membrane binding process is influenced by the initial C domain orientation, we simulated a series of systems (n= 8) where the wild type C2 domain was positioned 30 Å above the 10% DOPS lipids containing membrane. In these systems, the initial orientations of C2 domains were 0°, 45°, 90°, 135°, 180°, 225°, 270° and 315° respectively (α angle as defined in Figure 3) as shown in figure 3. During the approaching phase, the C2 domains underwent a series of rotations and translations. We observed no common mode for C2 domain movement during the approaching phase. Notably, the membrane-binding time for the C2 domain was similar (~1 microsecond) for all different initial orientations tested. We furthermore observed that at some time points during the approaching phase, several areas of the C2 domain touched the membrane surface but the domain did not anchor to the membrane (Figure 4B). The hydrophobic spikes 1 and 3 were always the first to anchor into the membrane regardless of the initial orientations of C2 domain. In all simulations, when the spikes1 and 3 were anchored, the C2 domain tilted, which facilitated the access of spike4 to the membrane surface.

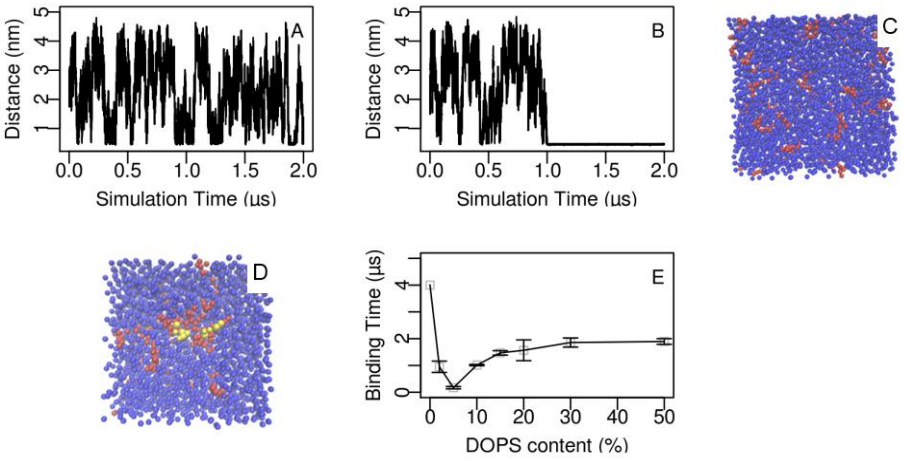


Figure 4. Influence of membrane composition in binding of the FVIII C2 domain. A. C2 domain binding in the absence of DOPS. B. C2 domain binding to DOPC membranes containing 10% DOPS. C. Top view on membrane, showing uniform distribution of DOPS (red) in DOPC (blue). D. Cluster formation of DOPS (red) when the C2 domain (The hydrophobic spikes shown in yellow) was bound to the membrane. E. membrane binding time of C2 domain in different DOPS contents in the membrane (0%, 2%, 5%, 10%, 15%, 20%, 30%, 50%). Shown are averages ± SD, n = 3.

The membrane and system's properties of the simulation systems Plasma membranes are mainly comprised by phospholipids that exist in all eukaryotes to maintain the integrity of the cell or organelles through their unique physiological functions such as semi-impermeable barrier, asymmetric distribution of phospholipids. The phospholipids are comprised of a polar headgroups (PC, PE, PS) and two hydrophobic tails (35, 39, 40). Following the martini model's parameters in terms of membrane quality, we used parameters that are within the allowed ranges, the density of the membrane surface should be in the range of 0.6-0.7 nm^2, the membrane thickness should be 4.83-5.13 nm, the lateral diffusion is 1e-7cm_2/sec (35, 41,

42) and the lateral order is -0.5 (anti-alignment) to 1 (perfect alignment) (41-44). After the 200 ns NPT simulation, the system's overall properties were evaluated including temperature, pressure and potential energy retrieved from the NPT trajectory from 150 ns to 200 ns. The properties that define the quality of a simulation system were calculated including the density, thickness, lateral diffusion and order of the lipids. All properties are consistent with the experimental data and within the acceptable ranges as defined by the martini force field (33, 35), see also Table 2.

Influence of membrane composition on C2 domain membrane binding

To test the influence of DOPS lipids on the FVIII C domain membrane binding process and to validate our molecular simulation model, we studied the binding time for C2 domain membrane binding at variable DOPS. As shown in Figure 4A, the C2 domain did not bind to the membrane in the absence of DOPS during the 2 μs simulation. Binding was not observed even when the simulation time was extended to 4 μs. When 10% DOPS was present in the membrane, the C2 domain bound to the membrane after around 1μs (Figure 4B). We observed that binding of the C2 domain influenced the distribution of DOPS lipids in the membrane layer contacting the C2 domain. In the absence of protein, DOPS is distributed evenly over the membrane layer (Figure 4C), but in the presence of the C2 domain, DOPS appeared to cluster in the membrane-binding area around the C2 hydrophobic spikes 1 and 3 of the C2 domain (Figure 4D). We have tested the membrane-binding time of C2 domain at

varying DOPS percentages (Figure 4E) and found that binding times decreased from 0 to 5% DOPS. At DOPS values >5%, the binding times prolonged slightly until an apparent plateau at 2 μs was reached.

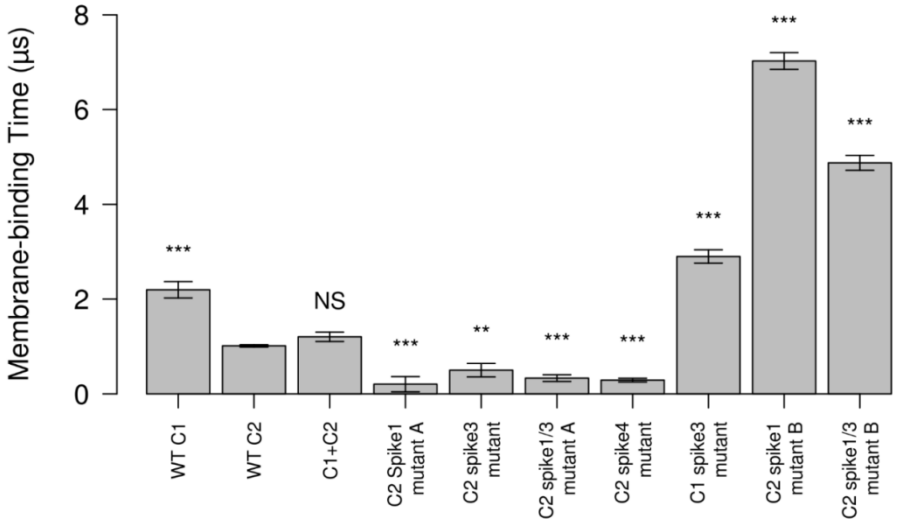


Figure 5. Membrane binding time for different FVIII C domain systems. All graphs show mean ± SD for 3 independent simulation experiments, statistical significance was tested using one way analysis of variance (ANOVA) with Dunnett post hoc test, * P<0.05, ** P<0.01, *** P<0.001. Significance in times was calculated by ANOVA and a star above the bar indicates significance. C2 domain and C1+C2 domain had a comparable binding time. The other systems were significantly different from the C2 system (significance calculated by ANOVA). C1, C1-spike3 mutant, C2-spike1 mutant B and C2-spike1/3 mutant B had about ~2, ~3, ~5, ~7 fold longer binding times respectively while C2-spike1 mutant A, C2-spike3 mutant, C2-spike1/3 mutant A and C2-spike4 mutant had faster membrane association abilities.

Initial membrane approach and binding of C domains

Comparison of membrane binding times (Figure 5), for 10% DOPS / 90% DOPC membranes, revealed that the binding time of C1+C2 was comparable to that of C2 alone, while C1 presented with an

almost 2-fold longer time than C2 or C1+C2. In the C1 domain system, we observed that spike 3 and 4, which contain bulky residues, bound to the membrane and anchored into the membrane. Next, the C1 domain tilted and residues located near spike 1 interacted with the membrane DOPS lipids. Interaction between Y2105 and K2110 and DOPS lipids caused the isolated C1 domain to tilt nearly 80° (α=79°) as shown in Figure 6A.

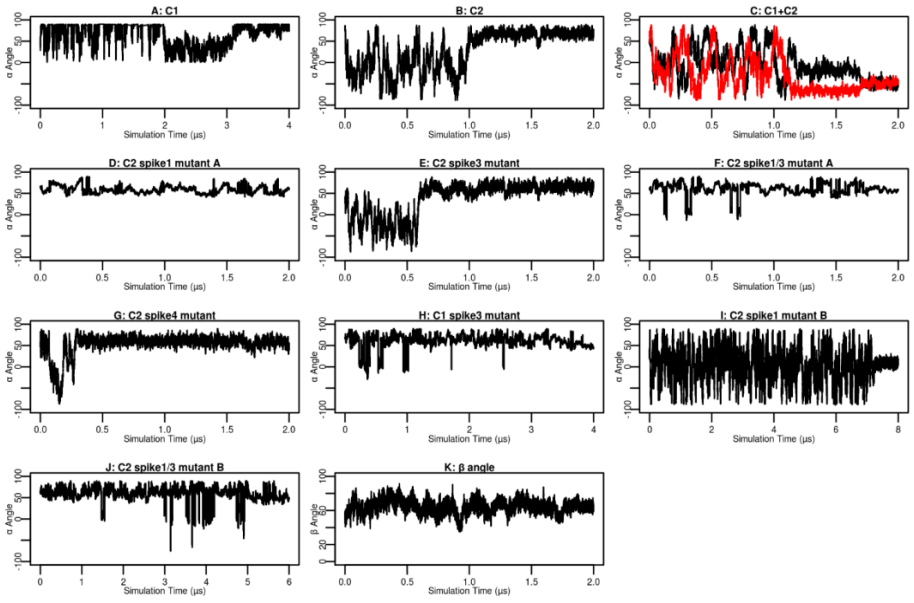


Figure 6. α, β angle movements during FVIII C domain membrane binding for different C domain variants tested. A. α angle for the wild type C1 system. B. α angle for the wild type C2 system. C. α angle for the wild type C1+C2 system (Black: C1 domain; Red: C2 domain). D. α angle in C2 spike1 mutant A. E. α angle in C2 spike3 mutant. F. α angle in C2 spike1/3 mutant A. G. α angle in C2 spike4 mutant. H. α angle in C1 spike3 mutant. I. α angle in C2 spike1/3 mutant B. J. α angle in C2 spike1/3 mutant B. K. β angle in C1+C2. The average angles from the last 50 snapshots are: wild type C1: 79°; wild type C2: 71°; C1 in C1+C2: -46°; C2 in C1+C2: -50°. C2 spike1 mutant A: 57°; C2 spike3 mutant: 70°; C2 spike1/3 mutant A: 58°; C2 spike4 mutant: 48°; C1 spike1 mutant B: 45°; C2 spike1/3 mutant B: 12°; C2 spike1/3 mutant B: 48°; β angle: 63°

It has to be noted that the interaction between these two residues (Y2105 and K2110) with the membrane was observed temporarily only in the isolated C1 domain. In the combined C1+C2 domain where the orientation of the C1 domain is restricted by the presence of the C2 domain these two residues cannot interact with the membrane. In the C2 domain system, the initial bound conformation had a tilting angle (α) of 24° with the region of A2201-S2206 touching the membrane and spike1 anchored into the membrane (Figure 6 and 7). The tilting angle (α) increased until spike3 and H2315 located at spike 4 were membrane-anchored. The isolated C2 domain anchoring in addition induced interaction between residues located near these 4 spikes (Y2195, T2197 and N2198, located nearby spike 1; R2215, V2223, N2224, K2227 located nearby spike 2; K2249 located nearby spike 3 and H2315) and the membrane. Finally, spike2 together with spike 3 were attracted by three DOPS lipids, which stably re-orientated the C2 domain conformation to 71° (α).

In the C1+C2 system, the C2 domain touched the membrane first, mediated through its four hydrophobic spikes at around 1 μs and next the C2 domain stabilized and anchored in membrane. This anchoring of C2 preceded the approach and binding of the C1 domain (see Figure 7C). Finally, the C1 and C2 were bound to the membrane at 1.7 μs with a corresponding increased β angle (as defined in Figure 3).

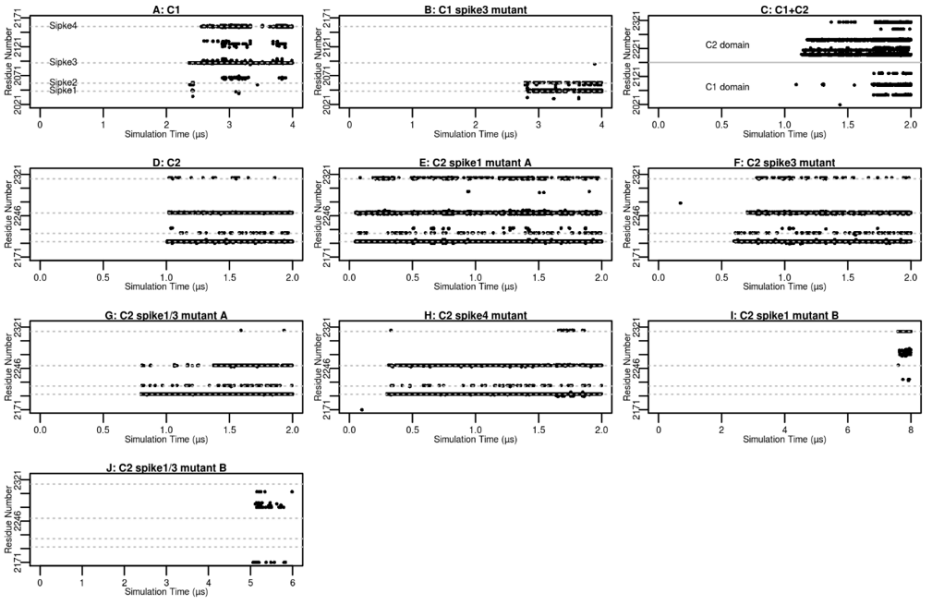


Figure 7. Membrane burial of FVIII C domains for each MD system tested. The x-axis indicates the simulation time while the y-axis indicates the residue numbering in the FVIII C domains, residues are indicated from N- to C-terminus. The four hydrophobic spikes were labeled and indicated by dotted lines for both C1 and C2 variants.

Three C2 domain mutants: C2 spike 1 mutant A, C2 spike 3 mutant and C2 spike 1/3 mutant A which have been described before (45), were analysed in order to further validate our molecular simulation model. For all of these 3 mutants, we confirmed, through independent computational method, their improved membrane binding, as can be estimated from the reduced binding times (Figure 5), that were previously determined for the corresponding variant proteins *in vitro* (45). Moreover, these mutants showed a similar binding mechanism as was observed for the wild type C2 domain (Figure 7). For mutants C1 spike 3 mutant, C2 spike 1/3 mutant B and C2 spike 1 mutant B, we observed a reduced membrane

binding, in agreement with functional characterization studies of the corresponding FVIII molecules (13, 25, 46, 47) thereby further supporting our approach.

We found that charged residues, such as K2065, K2072, K2110 and R2150 in the isolated C1 domain and R2215, R2220, K2249 in the isolated C2 domain (shown in Figure 8A and 8D, respectively) play an important role as electrostatic driving forces for the C domain membrane approach, prior to membrane binding. For the combined C1+C2 domain, key residues that play a crucial role for this process are mostly from the C2 domain as demonstrated in Figure 8C. Moreover, we found that the interaction per residue of K2249 in the C2 spike 4 mutant (Figure 8H) was increased as compared to the wild type C2 which causes the reduced binding time of this system. Our simulations demonstrated that electrostatic interactions play an important role during the approaching stage, which was further strengthened by the observation that mutation from neutral to positively charged residue (W2313R) in the C2 spike 4 mutant resulted in a reduced binding time, or on the other hand mutation from positively charged to neutral residue (K2092A/F2093A) in the C1 spike 3 mutant increased the binding time.

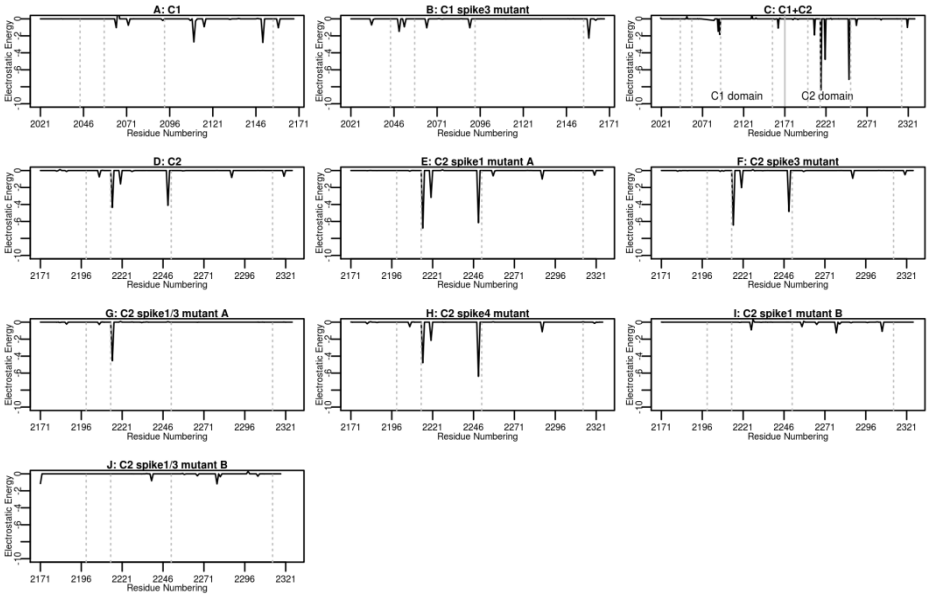


Figure 8. Electrostatic interaction (kJ/mol) between each FVIII C domain residue and the membrane during the approaching stage for different C domain variants. The X-axis indicates FVIII C domain residue numbering, the Y-axis indicates the electrostatic interaction which is calculated from the simulation trajectory in the approaching stage as described (48, 49). A, wild type C1. B, C1 spike3 mutant. C, wild type C1+C2 (domain boundary indicated). D, wild type C2. E, C2 spike1 mutant A. F, C2 spike3 mutant. G, C2 spike1/3 A mutant. H, C2 spike4 mutant. I, C2 spike1 mutant B. J, C2 spike1/3 B mutant.

Membrane-buried residues

To identify which amino acids are buried in the membrane during the membrane insertion phase, we analyzed the position of each individual amino acid side chain of the C1, the C2 and the C1+C2 domains during the simulations (Figure 7 and 9).

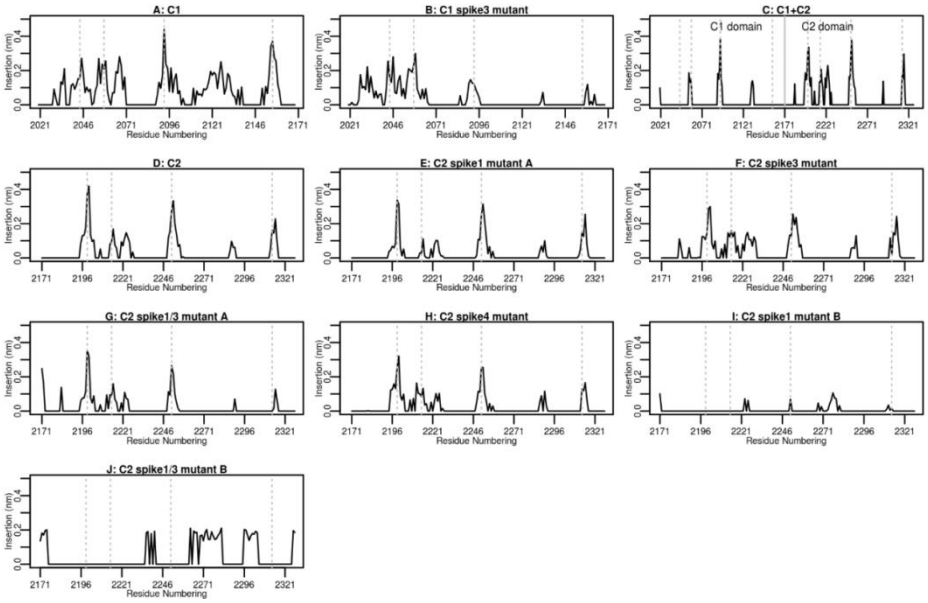


Figure 9. Depth of membrane-burial per each FVIII C domain. The X-axis indicates FVIII C domain residue numbering, the Y-axis indicates the distance between the anchored residues to the surface of the membrane (monitored by the NC3 group of DOPC lipids). The vertical dotted lines indicate the position of spikes 1-4 respectively. A. wild type C1 system. B. C1 spike3 mutant. 6C. C1+C2 (Divided by solid line). D. wild type C2. E. C2 spike1 mutant A. F. C2 spike3 mutant. G. C2 spike1/3 mutant A. H. C2 spike4 mutant. I. C2 spike1 mutant B. J. C2 spike1/3 B mutant.

In the C1 system (Figure 7A), spike1 and spike 2 did not substantially contribute to membrane insertion. On the other hand, spikes 3 and 4 and the intermediary residues (e.g. S2094, S2095, Y2097) were stably buried inside the membrane, with F2093 being dominantly buried in the membrane. In the FVIII C2 system (Figure 7D), we noted that in particular spikes 1 and 3 were inserted into the membrane.

When analysing the binding of the C1 + C2 domain pair (C1+C2), we observed a cooperative binding between the C1 and C2 domains

during membrane penetration (Figure 7C). Figure 7C shows that burial of the four C2 domain spikes at 1 μs precedes burial of all four C1 domain spikes at 1.7 μs. Binding of all four spikes of both the C1 and C2 in this combined system in addition induced interaction of F2290, W2313 and I2317, which are located near spike 4, with the membrane. The identified interactive residues in this study are thus well in agreement with reported experimental data which have attributed important functions to these spikes and individual amino acids(3, 4, 19, 21).

For 4 mutant C2 systems tested, a similar general pattern as for the wild type C2 domain was observed, yet binding occurred already at short binding times (cf. Figure 7 D-H). However, we found that spike 4 of the C2 spike 4 mutant and of C2 spike 1/3 mutant A cannot bury deeply into the membrane as shown in Figure 9. It was reported that mutants C1 spike 3, C2 spike 1/3 mutant B and C2 spike 1 mutant B can cause haemophilia A because the causative mutations presumably reduce the membrane binding ability (13, 25, 46, 47). Our simulations revealed (Figure 7, and 9) for these mutant C domains, that membrane binding as mediated by the C domain spikes deviates from that observed for the wild type domains, resulting in reduced membrane binding .

Binding Free Energy

We calculated the binding free energies for all simulation systems described here (Figure 10). C1 domain presented with a relatively low binding energy to the membrane of -35.39±0.42 kcal/mol, while

C2 binding was characterized by a -77.36±0.90 kcal/mol binding free energy. When C2 bound together with C1, the total binding free energy for the two domains binding to the membrane was -198.69±6.78 kcal/mol indicating the cooperative binding of the two domains, in agreement with experimental data (50), reporting that the C2 domain is of likely greater importance to membrane binding than the C1 domain, however the C1 domain is indispensable for overall binding.

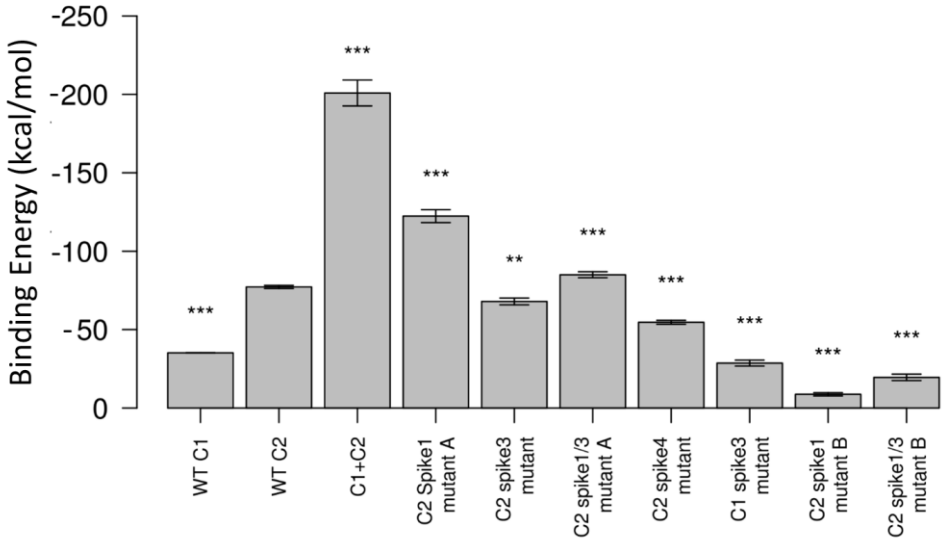


Figure 10. Membrane binding energy (ΔGbind) of FVIII C domains during membrane binding. Binding free energies for wild type and mutant domains are calculated by the potential mean force (PMF) as described (36, 37). The membrane contains 10% DOPS / 90% DOPC lipids. The binding energy for each system was compared to that of the wild type C2 domain. All graphs show mean ± SD, n=3 and statistical significance was tested using one way analysis of variance (ANOVA) with Dunnett post hoc test, * P<0.05, ** P<0.01, *** P<0.001.

The C2 spike 1 mutant A had an apparently ~2-fold higher binding energy than wild type C2, whereas average binding energies for C2

spike 3 mutant and C2 spike 1/3 mutant A were much comparable to that of the wild type C2 domain (Figure 10). These results are in agreement with an experimentally determined K_D for the wild type C2 domain of 0.46 ± 0.03 μM and a K_D for C2 spike 1 mutant A, C2 spike 3 mutant and C2 spike 1/3 mutant A of 0.19 ± 0.03 μM, 0.33 ± 0.03μM, and 0.19 ± 0.05 μM, respectively (45).

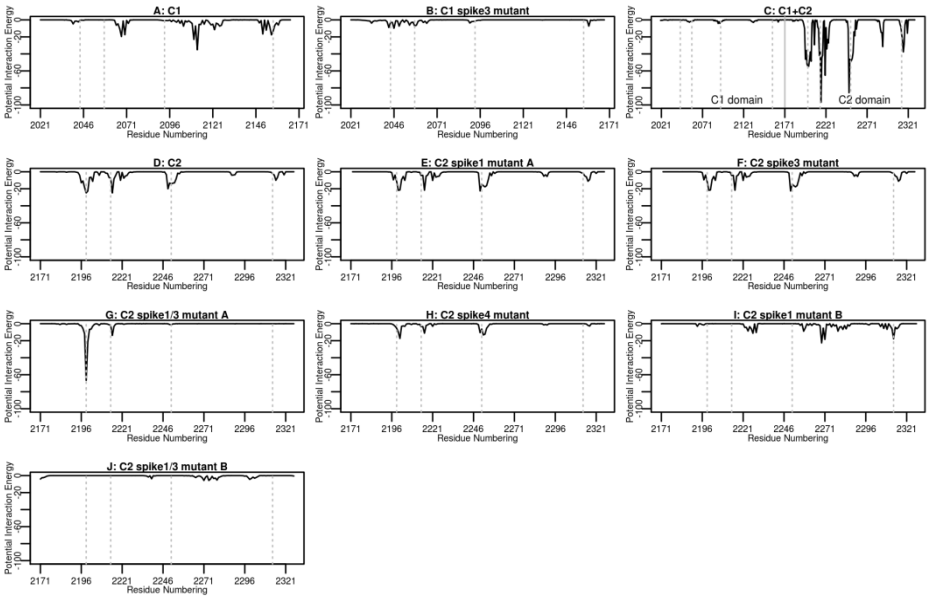


Figure 11. Binding free energy per residue in the membrane anchored stage. The X-axis indicates FVIII C domain residue numbering, and the Y-axis indicates membrane binding free energy in kJ/mol. A. wild type C1. B. C1 spike3 mutant. C. C1+C2 (Divided by solid line). D. wild type C2. E. C2 spike1 mutant A. F. C2 spike3 mutant. G. C2 spike1/3 mutant A. H. C2 spike4 mutant. I. C2 spike1 mutant B. J. C2 spike1/3 B mutant.

Our simulations further revealed that mutation at M2199 and F2200 in the C2 spike 1 mutant A and C2 spike 1/3 mutant A variants resulted in increased binding because the mutated spike 1 residues can bind to the membrane stronger than their wild type counterparts

as demonstrated in Figure 11D, E and G. Binding free energies of other mutant systems (C2 spike 4 mutant, C2 spike 1/3 mutant B, C2 spike 1 mutant B, and C1 spike 3 mutant) were significantly lower than for wild type C2 (Figure 10). These results indicate that these mutations can cause haemophilia A because these mutations reduce the efficiency of membrane binding.

During the simulated membrane binding of the C1, C2 or C1+C2 domains, three distinct phases could be discriminated:

1. approach phase; during which electrostatic attraction between the basic C domains and the net negatively charged membranes occurs. This phase ends when the minimum distance between the C domain and the membrane is stably less than 5 Å.

2. binding and anchoring phase; during binding, the hydrophobic spikes interact with the membrane via mainly van der Waals forces, hydrophobic residues are buried inside the membrane layer during this phase and a clustering of negatively charged lipid around the bound protein is observed. This phase ends when the when the tilting angle (α) of the C domain remains stable.

3. consolidation phase; during this phase conformational changes and C1+C2 domain reorientations occur which result in a stable binding conformation.

Discussion

We studied the influence of membrane composition on the

membrane binding process by CGMD simulation and used the isolated FVIII C2 domain as a model system, since it was reported before as the main membrane-binding motif (2), (12), (13). Our simulation results show that interaction with DOPS lipids is a prerequisite for FVIII C domain membrane binding. Residues at and around four hydrophobic spikes interact with DOPS lipids during the simulation, which is consistent with the general view on the absolute requirement for a net negatively charged membrane surface to support coagulation protein complexes (51). Importantly, the observation that DOPS is needed in our simulations represents a valuable validation of our simulation setup along with the data obtained for known FVIII mutants with altered membrane binding properties. We observed that membrane binding of the C2 domain influences the formation of DOPS lipid clusters as demonstrated in Figure 4D. Binding of the C2 domain thus appears an interplay between both lipid and protein: the negatively charged lipid is able to exert an attractive force on the protein that allows the translocation of protein from solution to the membrane surface, while, when bound, the protein induces a clustering of negatively charged lipids.

The membrane-binding time at varying DOPS content (Figure 4E) observed by us is in line with experimental results by Engelke et al. (52), who reported a similar pattern between the ratio of DOPS and FVIII's membrane-binding by calculation of the FVIII binding affinity (K_D). Since 10% of DOPS in the membrane, closely resembles the percentage of DOPS in an activated platelet (10%, (53)) we selected this lipid for our further simulation studies.

To access an optimal membrane binding interface, the C domains perform a series of rotations, translations, binding- and dissociation-interactions, while probing several different binding areas until the four hydrophobic spikes of the C domains bound initially to the membrane. As the FVIII C1 domain lacks bulky residues in its spike 1 and 2 that may serve to anchor these loops in the membrane, spikes 3 and 4 of the C1 domain play important roles as to initially tether the C1 to the membrane (Figure 7A). This is in contrast to binding of the C2 domain, where mainly spikes 1 and 3 mediate residue burial, with minor contributions from spikes 2 and 4 (Figure 7D). In case of the C1+C2 system, first the C2 domain bound followed by the C1 domain. The binding times of C domains (Figure 5) showed that C2 and C1+C2 bind to the membrane nearly at the same simulation time whereas the C1 domain requires more time. This result suggests that binding of the C2 domain is a dominant event in the FVIII C-domain mediated membrane binding process.

For the C2 mutants of known improved binding tested, we found that all four spikes bound to the membrane much like in the wild type C2 domain, yet at shorter binding times (Figure 5 and 4). Our simulations showed that mutation of K2092A/F2093A (C1 spike 3 mutant) and deletion of A2201 (C2 spike1 mutant B), which cause haemophilia A, resulted in reduced membrane binding ability (Figure 5 and 7). Moreover, we found that the tilting angles, α, of the C1 and C2 wild type and the mutant systems that cause haemophilia A are different (Figure 6), potentially adding to the loss in membrane binding. Our simulation results revealed that the optimal β angle

between C1 and C2 of the C1+C2 system in the membrane-bound phase was 63°, which is different from the unbound state of the crystallized FVIII structures (53°) (16-18). This difference may indicate that the crystallographic structure of FVIIIa, does not fully represent its membrane-bound conformation.

We identified which key residues contribute to the membrane interaction (Figure 12 and Table 3). It has been hypothesized that Q2213 and N2217 located at C2 spike 2 interact with a PLS head group (54). Our simulations illustrate that N2217 indeed interacts with the membrane but does not become inserted into the membrane, instead the neighboring residue Q2213 becomes buried in the membrane. Foster et al showed that when the region T2303 to Y2332 is blocked, the membrane binding ability of the C2 domain was lost by up to 90% (55). Results derived from our simulations confirm that residues from this same area, such as Q2311, W2313 (spike 4), V2314 (spike 4), H2315 (spike 4), R2320, are directly involved in the membrane binding process. In addition, spike 4 was found to be buried in the membrane during the simulation, which may explain why the compound 005B10, that binds to residues at spike 4 (W2313, V2314, H2315), can inhibit FVIII membrane-binding activity (4, 21). The involvement of W2313 in membrane binding, as discovered here, is a valuable addition to the explanation of the mild haemophilia A in carriers of the W2313R mutation. All the more since earlier attempts to recombinantly express this variant for functional studies, have failed (27), again showing added value of these simulation studies. Our simulation provides further molecular explanation as to why mutation W2313R can cause haemophilia A

because R2313 cannot bury deeply into membrane and thus the tilting angle α of this system is different from that of the wild type C2. This can result in a lower overall binding affinity for this mutant system as illustrated by the calculated binding free energy of this mutant (-54.65 ±1.28 kcal/mol) as compared to C2 wild type (-77.22 ± 1.01 kcal/mol) (Figure 10) (21). Moreover, the interaction energy per residue during the anchoring stage (Figure 11) showed that W2313 (-2.34 kcal/mol) can interact with the membrane stronger than the mutant R2313 (-0.12 kcal/mol).

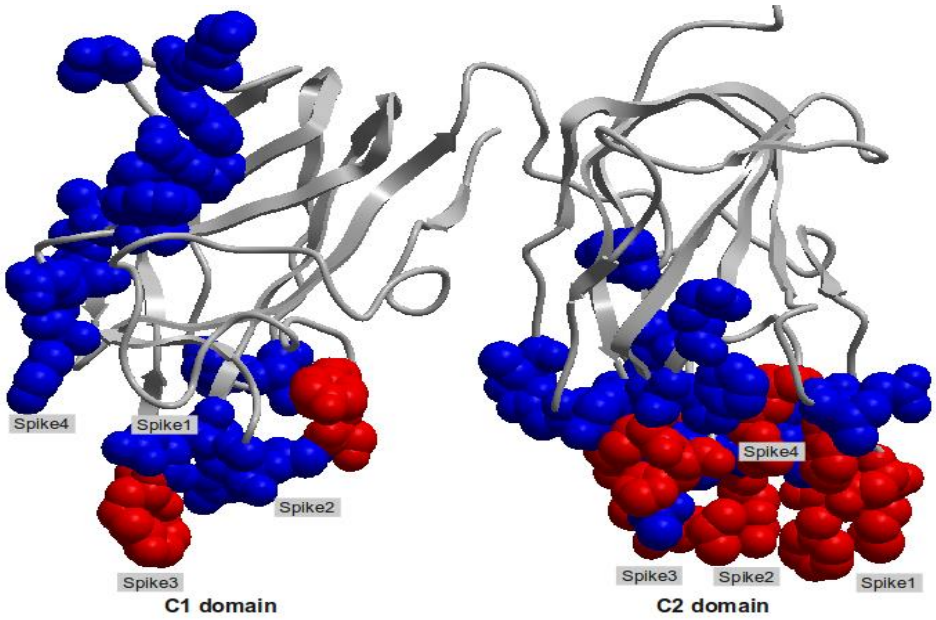
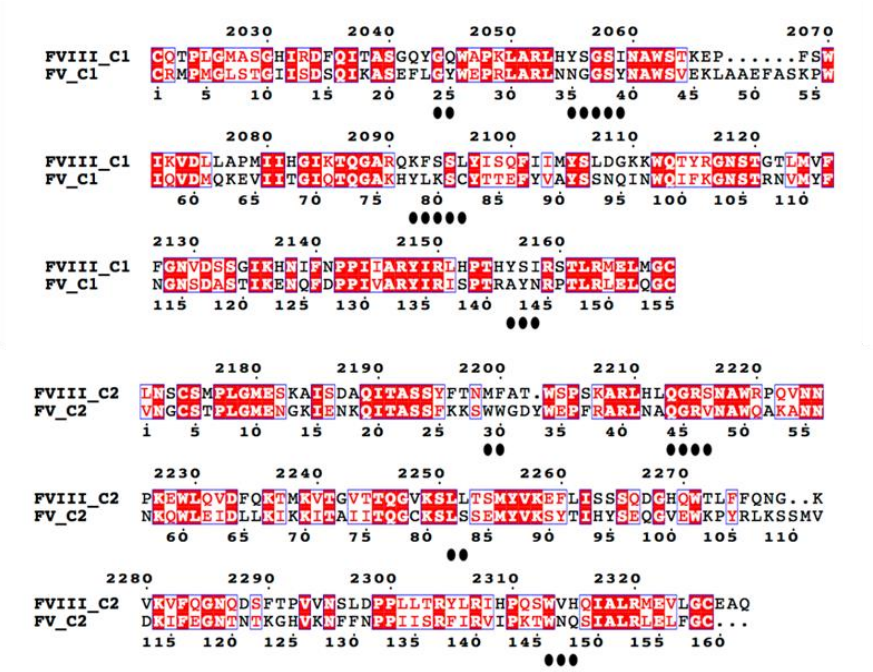


Figure 12. FVIII C domain residues involved in membrane-binding. C domain residues predicted to be involved in membrane binding either by contributing to the membrane approach, the interaction with the membrane or to membrane-anchoring, corresponding to the data in Table 3. The blue colored residues are confirmed haemophilia A associated while red colored residues are involved in membrane-binding but have not been identified in the context of haemophilia A.

The membrane-burial depth for each residue was studied (see Figure

7 and 9 for details) for wild-type and mutant C domains. In the wild type C1 domain, spikes1-4 and F2126 become inserted into the membrane, with spikes 3 and 4 penetrating deepest.



Domains	Spike1	Spike2	Spike3	Spike4
C1 FV	G1902-Y1903	N1913-Y1917	Y1956-C1960	A2020-N2022
C1 FVIII	G2044-Q2045	Y2055-I2059	K2092-L2096	Y2156-I2158
C2 FV	W2063-W2064	Q2078-N2082	L2116-S2117	W2180-Q2182
C2 FVIII	M2199-F2200	Q2213-S2216	L2251-M2255	W2313-H2315

Figure 13. Sequence similarity of the C domains between human FVIIIa and FVa. The sequence alignment was generated by the ESPript Web server. Human FVIII and FV share 48% (65%) sequence identity (similarity) in the C1 domain and 43% (65%) sequence identity (similarity) in the C2 domain. The FVIII numbering is indicated above the sequences and the scale is shown below. Identical or similar residues are indicated by blue boxes. The four conserved spikes are indicated by black circles and also defined in the table.

In the wild type C2 domain, residues from spike1 and 3 were deeper-buried than other residues, followed by the spike 4 and 2. The C2 mutants such as C2-spike 1 mutant A and C2-spike 3 mutant followed a similar pattern while in C2-spike 1/3 mutant A and C2-spike 4 mutant (W2313R), the burial depth of spike4 (W2313-H2315) was decreased. In the C1+C2 system, membrane burial resembled that for the isolated domains.

While we did not explicitly study the binding of FV C domains, given the high degree of structural and functional homology between FVIII and FV, and the presence of paralogous spikes 1-4 in FV (Figure 13) we hypothesize that also the binding process of FV might follow an analogous path as observed here for FVIII. This assumption is supported by the observations made by us for C2-spike 1 mutant A, C2-spike 1/3 mutant and C2-spike 3 where in fact spikes in FVIII have been replaced by the homologous spikes from FV. Whereas these variants showed improved binding in terms of the time required for binding (Figures 6 and 7), the binding pattern was in fact similar to that of the WT FVIII C2. Thus, the membrane binding mechanism observed here by us for the C domains of FVIII could represent a conserved mode of calcium independent membrane binding. A separate experiment in which we have simulated the binding of the FV C2 domain confirms our hypothesis.

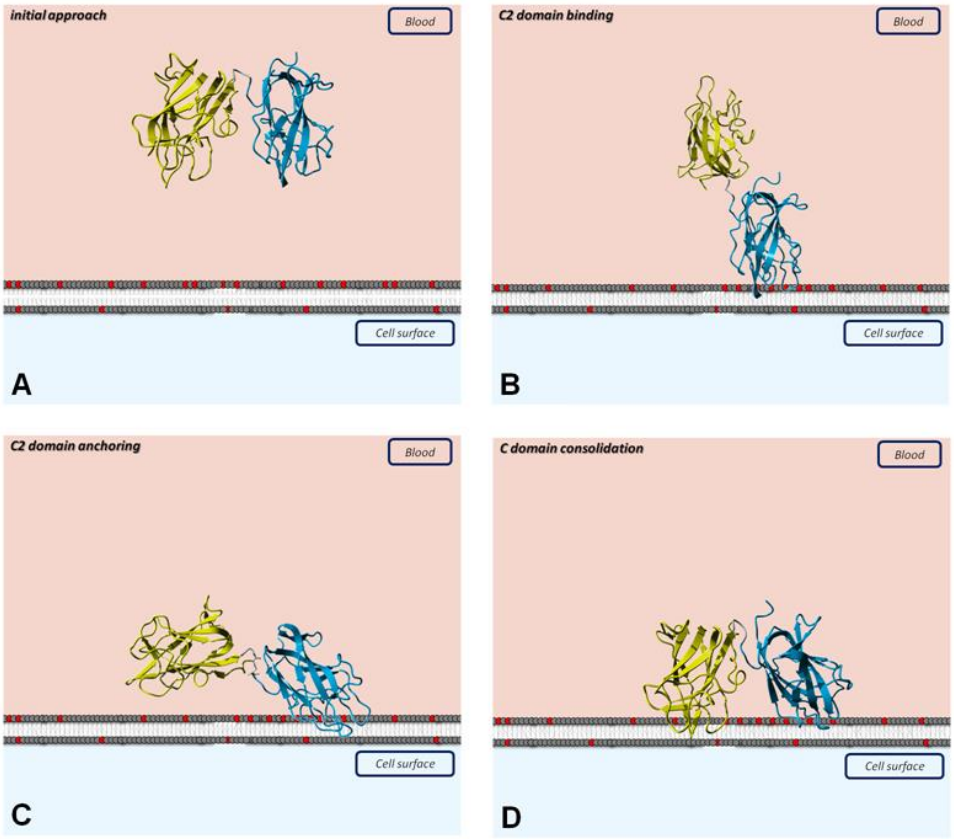


Figure 14. General membrane binding mechanism of the FVIIIa C domains. Simplified cartoons illustrate the different steps involved in membrane binding, as observed in this study. A. Initial approach of the C1+C2 domains to the membrane by electrostatic interaction between the net positively charged C domains and a negatively charged PS exposing cell surface (the latter indicated as red dots). B. Anchoring of the C2 domain in the DOPS containing membrane, which increases the angle between the C1 and C2 domain. C. Rearrangement of orientation of the C2 domain towards the membrane facilitates the approaching of the C1 domain to the membrane surface while spike interaction will anchor both domains in the membrane. D. Both the C2 and the C1 domain are anchored into the membrane resulting in stable binding of the C1 and C2 domains, with the C1 and C2 domains at non-perpendicular angles to the membrane.

615 FVIII missense mutations have been reported to cause haemophilia A (HAMSTeR database, Feb 2013) while causal relationships between haemophilia A and missense mutations are

variable (24, 45, 56). We have listed C domain residues which are involved in haemophilia A and identified residues that are important for membrane binding of the combined FVIII C1 and C2 domains in Table 3. Mutation of several FVIII C domain residues can result in defective phospholipid binding. As illustrated in Figure 12, we have correctly identified through *in silico* simulation, residues which are involved in membrane binding and which are related to haemophilia A. Moreover, additional residues which participate in the membrane binding process but which have not been reported earlier in haemophilia A were identified and we propose that these residues are prime candidates for novel haemophilia A causing mutations. Based on our simulations we can provide information about individual residues that cause haemophilia A. For example, we identified that the presence of A2201 is important for membrane binding. Further, it was reported (57) that mutation of R2150H results in impaired FVIII binding to VWF and thus can cause mild-moderate haemophilia (57). Our simulations indicate that residue 2150 is important for membrane binding as well (Table 3). Mutation of Q2311P can cause severe haemophilia A but the functional cause was not yet identified (57). Our results now reveal that Q2311 is involved in membrane binding.

In conclusion, we have applied CGMD to study the binding mechanism of the C1, C2 and C1+C2 domain of FVIII to phospholipid membranes. From our simulations we propose a general mechanism for the combined binding of the C1 and C2 domain to a membrane (Figure 14), where the C2 domain will bind first and becomes anchored to the membrane surface. Such anchoring is accomplished by penetration of hydrophobic spikes into the lipid layer. Next, the C2

domain facilitates a C1 membrane binding interaction which results in a conformational change of the C1+C2 dimer: the angle between the C1+C2 domains changes as compared to the unbound state, ultimately resulting in a bound C1+C2 dimer.

Our results indicate that binding of the C2 domain is a main driving event for the binding of the C1+C2 dimer to the membrane. Moreover, we have found that electrostatic interactions are the main driving force to drive the C domain towards the membrane and we identified the residues involved in FVIII C-domain mediated membrane binding. These data provide a likely rationale that can help to explain phenotypes observed in haemophilia A, both for known mutations as well for yet undiscovered C-domain mutations.

Table 1. Human FVIII C domain mutants. 7 mutants which have been *in-silico* studied are described with their names, location and the desired residues to be mutated.

Mutant Name	C1-Spike 3	C2-Spike1	C2-Spike 3	C2-Spike 4	Function and Properties	Ref.
C2-spike 1 mutant A		M2199W F2200W			Improved binding	(45)
C2-spike 1/3 mutant A		M2199W F2200W	L2252S		Improved binding	(45)
C2-spike 3 mutant			L2252S		Improved binding	(45)
C2-spike 1/3 mutant B		M2199A; F2200A	L2251A L2252A		Reduced binding	(13)
C2-spike 1 mutant B		deletion of A2201			Reduced binding	(25)
C1-spike mutant 3	K2092A; F2093A				Reduced binding	(46)
C2-spike mutant 4				W2313R	Mutation cannot be made experimentally	(27, 58)

Table 2. The table presents the parameters that describe the quality of the molecular dynamics simulation systems used in this study.

	C1 system	C2 system	C1+C2 system
Nrs DOPC/DOPS	178/20	240/24	352/38
Temperature (K)	324.31	324.57	324.54
Area of lipid (nm2/lipid)	0.697	0.700	0.69 9
Thickness (nm)	4.97	4.946	4.96
Lateral diffuse (1e-7cm2/sec)	3.29	6.0	6.92
Alignment order	0.315	0.300	0.31

Table 3. Residues predicted to be involved in membrane-binding are presented and categorized according to the type of contribution they provide to the overall membrane binding process. Approaching Force: these residues contribute by means of electrostatic interaction; Interaction: these residues contribute by means of electrostatic and van der Waals forces, and Membrane-buried: these residues become buried in the lipid layer and act as anchors. (23). We arbitrarily set the threshold for approaching interactions at < -0.7 kJ/mol as "Y", and all energies >-0.7 kJ/mol as "N", for the interaction the threshold was set at -20 kJ/mol and the membrane burial threshold was set at 5 Å. Resides that are known to contribute to haemophilia A are indicated in the last column.

Residues number	electrostatics interaction	interaction energy	membrane buried >5A	Haemophilia A
C1 domain				
W2046	N	N	Y	Y
Y2055	N	N	Y	N
W2062	N	N	Y	Y
K2065	Y	N	N	Y
E2066	N	Y	N	Y
P2067	N	Y	N	N
F2068	N	Y	Y	N
W2070	N	Y	Y	Y
K2072	Y	N	N	N
R2090	N	N	Y	Y
K2092	Y	N	Y	N
F2093	N	N	Y	N
Y2097	N	N	Y	N
D2108	Y	N	N	N
K2110	Y	Y	Y	Y
K2111	Y	N	N	N
W2112	N	Y	Y	Y
R2116	Y	N	N	Y
T2122	N	Y	N	Y
Y2148	N	N	Y	N
R2150	Y	Y	N	Y
H2152	N	Y	N	Y
H2155	N	Y	Y	Y
Y2156	N	Y	Y	N
R2159	Y	N	N	Y
C2 domain				
F2196	N	Y	Y	N
T2197	N	Y	N	N
N2198	N	Y	Y	N
M2199	N	Y	Y	N
F2200	N	Y	Y	N
A2201	N	N	Y	Y
W2203	N	Y	Y	Y
K2207	Y	N	N	N
Q2213	N	Y	N	Y
R2215	Y	Y	Y	N
N2217	N	Y	Y	N
R2220	Y	Y	N	N
K2249	Y	Y	Y	N

S2250	N	Y	Y	N
L2251	N	Y	Y	N
L2252	N	Y	Y	N
T2253	N	Y	Y	Y
S2254	N	N	Y	N
Q2311	N	Y	Y	Y
W2313	N	Y	Y	Y
V2314	N	Y	Y	N
H2315	N	Y	Y	N
R2320	Y	N	N	Y

C1+2 domain pair				
Y2055	N	N	Y	N
R2090	Y	N	N	Y
K2092	Y	N	N	N
F2093	N	N	Y	N
R2163	Y	N	N	Y
F2196	N	Y	N	N
T2197	N	Y	N	N
N2198	N	Y	N	N
M2199	N	Y	Y	N
F2200	N	Y	Y	N
A2201	N	Y	N	Y
T2202	N	Y	N	N
W2203	N	Y	N	Y
K2207	Y	Y	N	N
Q2213	N	Y	N	Y
G2214	N	Y	N	N
R2215	Y	Y	Y	N
S2216	N	Y	N	N
N2217	N	Y	N	N
R2220	Y	N	N	N
Q2222	N	Y	N	Y
V2223	N	Y	N	Y
N2224	N	Y	N	N
K2249	Y	Y	N	N
S2250	N	Y	N	N
L2251	N	Y	Y	N
L2252	N	Y	Y	N
T2253	N	Y	N	Y
S2254	N	Y	N	N
Y2256	N	N	N	Y
K2258	Y	N	N	N
D2288	N	N	N	Y
F2290	N	Y	N	N
W2313	N	N	N	Y
V2314	N	Y	N	N
H2315	N	Y	N	N
Q2316	N	Y	N	N
R2320	Y	N	N	Y

Acknowledgments

We thank the Bayer Haemophilia Awards program for their support with a special project grant to G.A.F.N. and the China Scholarship Council (Grant No. 2008630114 to J.D.).

References

1. Lenting PJ, van Mourik JA, & Mertens K (1998) The life cycle of coagulation factor VIII in view of its structure and function. *Blood* 92(11):3983-3996.
2. Saenko E, *et al.* (2001) Comparison of the properties of phospholipid surfaces formed on HPA and L1 biosensor chips for the binding of the coagulation factor VIII. *J Chromatogr A* 921(1):49-56.
3. Novakovic VA, *et al.* (2011) Membrane-binding properties of the Factor VIII C2 domain. *Biochem J* 435(1):187-196.
4. Liu Z, *et al.* (2010) Trp2313-His2315 of factor VIII C2 domain is involved in membrane binding: structure of a complex between the C2 domain and an inhibitor of membrane binding. *J Biol Chem* 285(12):8824-8829.
5. Lu J, Pipe SW, Miao H, Jacquemin M, & Gilbert GE (2011) A membrane-interactive surface on the factor VIII C1 domain cooperates with the C2 domain for cofactor function. *Blood* 117(11):3181-3189.
6. Takeshima K, Smith C, Tait J, & Fujikawa K (2003) The preparation and phospholipid binding property of the C2 domain of human factor VIII. *Thromb Haemost* 89(5):788-794.
7. Pratt KP, *et al.* (1999) Structure of the C2 domain of human factor VIII at 1.5 A resolution. *Nature* 402(6760):439-442.
8. Eaton D, Rodriguez H, & Vehar GA (1986) Proteolytic processing of human factor VIII. Correlation of specific cleavages by thrombin, factor Xa, and activated protein C with activation and inactivation of factor VIII coagulant activity. *Biochemistry* 25(2):505-512.
9. van Dieijen G, Tans G, Rosing J, & Hemker HC (1981) The role of phospholipid and factor VIIIa in the activation of bovine factor X. *J Biol Chem* 256(7):3433-3442.
10. van Dieijen G, van Rijn JL, Govers-Riemslag JW, Hemker HC, & Rosing J (1985) Assembly of the intrinsic factor X activating complex--interactions between factor IXa, factor VIIIa and phospholipid. *Thromb Haemost* 53(3):396-400.
11. Gilbert GE & Arena AA (1996) Activation of the factor VIIIa-factor IXa enzyme complex of blood coagulation by membranes containing phosphatidyl-L-serine. *J Biol Chem* 271(19):11120-11125.
12. Foster PA, Fulcher CA, Houghten RA, & Zimmerman TS (1990) A synthetic factor VIII peptide of eight amino acid residues (1677-1684) contains the binding region of an anti-factor VIII antibody which inhibits the binding of factor VIII to von Willebrand factor. *Thromb Haemost* 63(3):403-406.
13. Gilbert GE, Kaufman RJ, Arena AA, Miao H, & Pipe SW (2002) Four hydrophobic amino acids of the factor VIII C2 domain are constituents of both the membrane-binding and von Willebrand factor-binding motifs. *J Biol Chem* 277(8):6374-6381.

14. Stace CL & Ktistakis NT (2006) Phosphatidic acid- and phosphatidylserine-binding proteins. *Biochim Biophys Acta* 1761(8):913-926.
15. Saenko EL, Ananyeva NM, Tuddenham EG, & Kemball-Cook G (2002) Factor VIII - novel insights into form and function. *Br J Haematol* 119(2):323-331.
16. Ngo JC, Huang M, Roth DA, Furie BC, & Furie B (2008) Crystal structure of human factor VIII: implications for the formation of the factor IXa-factor VIIIa complex. *Structure* 16(4):597-606.
17. Shen BW, *et al.* (2008) The tertiary structure and domain organization of coagulation factor VIII. *Blood* 111(3):1240-1247.
18. Svensson LA, Thim L, Olsen OH, & Nicolaisen EM (2013) Evaluation of the metal binding sites in a recombinant coagulation factor VIII identifies two sites with unique metal binding properties. *Biol Chem* 394(6):761-765.
19. Stoilova-McPhie S, Lynch GC, Ludtke S, & Pettitt BM (2013) Domain organization of membrane-bound factor VIII. *Biopolymers* 99(7):448-459.
20. Autin L, *et al.* (2005) Molecular models of the procoagulant factor VIIIa-factor IXa complex. *J Thromb Haemost* 3(9):2044-2056.
21. Stoilova-McPhie S, Villoutreix BO, Mertens K, Kemball-Cook G, & Holzenburg A (2002) 3-Dimensional structure of membrane-bound coagulation factor VIII: modeling of the factor VIII heterodimer within a 3-dimensional density map derived by electron crystallography. *Blood* 99(4):1215-1223.
22. Nicolaes GA, Villoutreix BO, & Dahlback B (2000) Mutations in a potential phospholipid binding loop in the C2 domain of factor V affecting the assembly of the prothrombinase complex. *Blood coagulation & fibrinolysis : an international journal in haemostasis and thrombosis* 11(1):89-100.
23. Kemball-Cook G, Tuddenham EG, & Wacey AI (1998) The factor VIII Structure and Mutation Resource Site: HAMSTeRS version 4. *Nucleic Acids Res* 26(1):216-219.
24. Kim SW, *et al.* (2000) Identification of functionally important amino acid residues within the C2-domain of human factor V using alanine-scanning mutagenesis. *Biochemistry* 39(8):1951-1958.
25. d'Oiron R, *et al.* (2004) Deletion of alanine 2201 in the FVIII C2 domain results in mild hemophilia A by impairing FVIII binding to VWF and phospholipids and destroys a major FVIII antigenic determinant involved in inhibitor development. *Blood* 103(1):155-157.
26. Tagariello G, *et al.* (2000) Experience of a single Italian center in genetic counseling for hemophilia: from linkage analysis to molecular diagnosis. *Haematologica* 85(5):525-529.
27. Spiegel PC, Murphy P, & Stoddard BL (2004) Surface-exposed hemophilic mutations across the factor VIII C2 domain have variable effects on stability and binding activities. *J Biol Chem* 279(51):53691-53698.
28. Khalid S & Bond PJ (2013) Multiscale molecular dynamics simulations of membrane proteins. *Methods Mol Biol* 924:635-657.
29. Krieger E, Koraimann G, & Vriend G (2002) Increasing the precision of comparative models with YASARA NOVA--a self-parameterizing force field. *Proteins* 47(3):393-402.
30. Joosten RP, *et al.* (2011) A series of PDB related databases for everyday needs. *Nucleic Acids Res* 39(Database issue):D411-419.
31. Kabsch W & Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*

22(12):2577-2637.
32. X. Periole MC, S.J. Marrink, M. Ceruso (2009) Combining an elastic network with a coarse-grained molecular force field: structure, dynamics and intermolecular recognition. *J. Chem. Th. Comp.* (5):2531-2543.
33. Marrink SJ, Risselada HJ, Yefimov S, Tieleman DP, & de Vries AH (2007) The MARTINI force field: coarse grained model for biomolecular simulations. *J Phys Chem B* 111(27):7812-7824.
34. Lee H & Larson RG (2008) Coarse-grained molecular dynamics studies of the concentration and size dependence of fifth- and seventh-generation PAMAM dendrimers on pore formation in DMPC bilayer. *J Phys Chem B* 112(26):7778-7784.
35. Anfinsen CB (1972) The formation and stabilization of protein structure. *Biochem J* 128(4):737-749.
36. Yesylevskyy SO, Schafer LV, Sengupta D, & Marrink SJ (2010) Polarizable water model for the coarse-grained MARTINI force field. *PLoS computational biology* 6(6):e1000810.
37. Djurre H. de Jong GS, W. F. Drew Bennett, Clement Arnarez, Tsjerk A. Wassenaar, Lars V. Schäfer, Xavier Periole, D. Peter Tieleman, and Siewert J. Marrink (2013) Improved Parameters for the Martini Coarse-Grained Protein Force Field. *J. Chem. Theory Comput.* 9 (1): 687–697.
38. Shankar Kumar JMR, Djamal Bouzida, Robert H. Swendsen, Peter A. Kollman (1992) THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *Journal of Computational Chemistry* 13(8):1011-1021.
39. van Meer G, Voelker DR, & Feigenson GW (2008) Membrane lipids: where they are and how they behave. *Nature reviews. Molecular cell biology* 9(2):112-124.
40. Fadeel B & Xue D (2009) The ins and outs of phospholipid asymmetry in the plasma membrane: roles in health and disease. *Crit Rev Biochem Mol Biol* 44(5):264-277.
41. Gaede HC & Gawrisch K (2003) Lateral diffusion rates of lipid, water, and a hydrophobic drug in a multilamellar liposome. *Biophysical journal* 85(3):1734-1740.
42. Balgavy P, *et al.* (2001) Bilayer thickness and lipid interface area in unilamellar extruded 1,2-diacylphosphatidylcholine liposomes: a small-angle neutron scattering study. *Biochim Biophys Acta* 1512(1):40-52.
43. Bulacu M, Periole X, & Marrink SJ (2012) In silico design of robust bolalipid membranes. *Biomacromolecules* 13(1):196-205.
44. Kurad D, Jeschke G, & Marsh D (2004) Lateral ordering of lipid chains in cholesterol-containing membranes: high-field spin-label EPR. *Biophysical journal* 86(1 Pt 1):264-271.
45. Gilbert GE, Novakovic VA, Kaufman RJ, Miao H, & Pipe SW (2012) Conservative mutations in the C2 domains of factor VIII and factor V alter phospholipid binding and cofactor activity. *Blood* 120(9):1923-1932.
46. Meems H, Meijer AB, Cullinan DB, Mertens K, & Gilbert GE (2009) Factor VIII C1 domain residues Lys 2092 and Phe 2093 contribute to membrane binding and cofactor activity. *Blood* 114(18):3938-3946.
47. Ettinger RA, James EA, Kwok WW, Thompson AR, & Pratt KP (2010) HLA-DR-restricted T-cell responses to factor VIII epitopes in a mild haemophilia A family with missense substitution A2201P. *Haemophilia : the official journal of the*

World Federation of Hemophilia 16(102):44-55.
48. Pronk S, *et al.* (2013) GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* 29(7):845-854.
49. May A, *et al.* (2014) Coarse-grained versus atomistic simulations: realistic interaction free energies for real proteins. *Bioinformatics* 30(3):326-334.
50. Wakabayashi H, Griffiths AE, & Fay PJ (2010) Factor VIII lacking the C2 domain retains cofactor activity in vitro. *J Biol Chem* 285(33):25176-25184.
51. Zwaal RF, Comfurius P, & van Deenen LL (1977) Membrane asymmetry and blood coagulation. *Nature* 268(5618):358-360.
52. Engelke H, Lippok S, Dorn I, Netz RR, & Radler JO (2011) FVIII binding to PS membranes differs in the activated and non-activated form and can be shielded by annexin A5. *J Phys Chem B* 115(44):12963-12970.
53. Heemskerk JW, Bevers EM, & Lindhout T (2002) Platelet activation and blood coagulation. *Thromb Haemost* 88(2):186-193.
54. Gilbert GE & Drinkwater D (1993) Specific membrane binding of factor VIII is mediated by O-phospho-L-serine, a moiety of phosphatidylserine. *Biochemistry* 32(37):9577-9585.
55. Foster PA, Fulcher CA, Houghten RA, & Zimmerman TS (1990) Synthetic factor VIII peptides with amino acid sequences contained within the C2 domain of factor VIII inhibit factor VIII binding to phosphatidylserine. *Blood* 75(10):1999-2004.
56. d'Oiron R, *et al.* (2006) Impact of choice of treatment for bleeding episodes on inhibitor outcome in patients with mild/moderate hemophilia a and inhibitors. *Semin Hematol* 43(1 Suppl 1):S3-9.
57. Graw J, *et al.* (2005) Haemophilia A: from mutation analysis to new therapies. *Nature reviews. Genetics* 6(6):488-501.
58. Schatz SM, *et al.* (2004) Mutation of the surface-exposed amino acid Trp to Ala in the FVIII C2 domain results in defective secretion of the otherwise functional protein. *Br J Haematol* 125(5):629-637.

Chapter 4

The structure function of the death domain of human IRAK-M

Jiangfeng Du, Gerry A.F. Nicolaes, Danielle Kruijswijk, Tom van der Poll, Miranda Versloot and Cornelis van ’t Veer

Submitted to Cell Communication and Signaling

Abstract

IRAK-M is an IRAK-4 dependent inhibitor of Toll-like receptor signaling in monocytes and macrophages. Lack of IRAK-M in mice greatly improves the resistance to nosocomial pneumonia and lung tumors. The rational design of IRAK-M inhibitors, which could potentially improve innate immunity in vulnerable patients with immunoparalysis, has thus far been impossible due to a virtual absence of detailed structure-function knowledge of IRAK-M. Since N-terminal death domain's (DD's) provide the primary interactions of IRAK molecules we generated a 3D structure model of the human IRAK-M-DD, to guide mutagenesis studies and predict protein-protein interaction points. Characterization of IRAK-M variant molecules indicated that both W74 and R97 in the DD are important for the NF-κB and ERK activating activity of IRAK-M as well as for the inhibitory action of IRAK-M on TLR induced release of cytokines. Furthermore, residues R70 and D19-A23 are specifically involved in ERK activation and protein levels of IRAK-M. W74 and R97 are located on opposite sides of the IRAK-M-DD and we hypothesize that IRAK-4-DD's will be sandwiched by R97-IRAK-M-DD and W74-IRAK-M-DD type interactions.

Keywords IRAK-M, inflammation, Death Domain, structure-function, TLR

Introduction

Interleukin-1 receptor-associated kinase M (IRAK-M) is a member of the IRAK protein family that is crucially involved in signaling initiated by IL-1, IL-18 or Toll-like receptor activation (1, 2). Activation of the receptors leads to dimerization of the adaptor MyD88 and subsequent recruitment of IRAK-4 and other IRAK's to form multimers (myddosomes) by homo- and heteromeric interactions of the death domains (3, 4). Binding and phosphorylation events triggered by IRAK-4 result in hyper- and auto-phosphorylation of IRAK-1 and formation of IRAK-1/TRAF-6 complexes which dissociate from the receptor to activate TAB2/3 and TAK-1 (5). TAB/TAK/TRAF6 activity leads to IκBα phosphorylation and ubiquitination culminating in nuclear factor-kB (NF-κB) activation and transcription of inflammatory genes (5). IRAK-2 hyperphosphorylation and TAB/TAK/TRAF6 activity leads to specific IRAK-2 dependent mRNA stabilization and translational control of pro-inflammatory mediators (6-8). Structurally, IRAK-M consists of a kinase domain (KD) flanked by an N-terminal death domain (DD) involved in binding to other IRAK family members and an unstructured C-terminal domain (CTD) with a TRAF6 binding motif. Just recently it was shown for murine IRAK-M that it is redundant with IRAK-1/2 in respect to NF-kB activation by a unique IRAK-4/IRAK-M mediated MEKK3 activation pathway (9). IRAK-1 and IRAK-4 contain active kinase subunits, in contrast IRAK-M and IRAK-2 lack the critical active site aspartate residue and are devoid of kinase activity (1, 10). All IRAK family members, mediate activation of NF-κB and MAPK (1) the phenotype of IRAK-1, IRAK-2 and IRAK-4 deficient mice or cells is one of

decreased production of inflammatory mediators (5). In contrast, IRAK-M deficient mice or cells display an increased inflammatory response (10). IRAK-M expression is induced upon TLR stimulation and IRAK-M inhibits cytokine and chemokine expression (9, 11). At first it was hypothesized that IRAK-M functioned by stabilizing the IRAK-1/IRAK-4 complex with MyD88 thereby preventing formation of IRAK-1/TRAF6 complexes (11), however other mechanisms have been put forward by which IRAK-M may inhibit inflammation in a more active manner. Among these are IRAK-M dependent stabilization of MKP-1 (12), down-regulation of the non-canonical NF-κB pathway (13), the specific induction of other negative regulators that are not regulated by mRNA stabilization such as A20, IκBα, SOCS-1 and SHIP by IRAK-M (9), and inhibition of IRAK-2 dependent mRNA stabilization/translation of cytokines and chemokines (9). Increased host responses caused by lack of IRAK-M are favorable for outcome in bacterial pneumonia (14-16) but also in tumor models (17) and bone marrow transplantation (18) which implies that inhibition of IRAK-M might have therapeutic potential. Different from the other IRAK's, IRAK-M expression is rather restricted to certain cell types such as monocytes/macrophage and lung epithelial cells (1, 19) where it is up-regulated under inflammatory conditions (20).

The W74 residue in the death domain of murine IRAK-M was shown to be crucial for the interaction with IRAK-4 and NF-κB activation (9). In this study we investigated the structure-function relationships of the death domain of human IRAK-M based on unbiased prediction of protein interaction sites and mutagenesis. We found that there are

two IRAK-4 binding sites involved in the NF-κB activating activity of IRAK-M with W74 and R97 as crucial residues which are both important for the inhibition of cytokine production by IRAK-M in monocytes. An area located in between W74 and R97 appears to be involved in another type of inhibition which is displayed in TLR mediated chemokine production by lung epithelial cells. The latter inhibitory action is provided by the stretch D19-A23 and R70 which is predicted to be involved in IRAK-M/IRAK-2 interaction. Intriguingly, neither W74 nor R97 appeared to be involved in this IRAK-M function in lung epithelium. Thus, we generated a high quality structure model of the death domain of IRAK-M that enables guided structure–function studies of human IRAK-M.

Results

Homology model of the human IRAK-M death domain

We generated a model for the death domain of human IRAK-M (IRAK-M-DD) by homology modeling based on the crystal structure of the death domain of mouse IRAK-4 (PDB 2A9I (21), which has 28.7% sequence identity to the human IRAK-M DD) as described in the Methods section. The generated IRAK-M-DD structure (Fig.1A) with 6 helical bundles forms a hydrophobic core that is decorated with a charged outer layer. An anti-parallel beta sheet, not seen in the template structure is formed by one strand from the N-terminus and another strand N-terminal of helix 5. An anti-parallel sheet located in between helix2 and helix3 in the template structure is absent in the DD of IRAK-M, instead a beta turn is made here by two serines in our

model. Unconstrained molecular dynamics simulation for 100 nanoseconds (ns) indicated good stability of this structure (Fig.2) and the quality of the structure was further verified by means of the total energy, root mean square deviation (RMSD) and the number of hydrogen bonds in the DD domain. Residues predicted to be involved in protein-protein interactions were identified as described in the Methods section. The identified interactive residues are gathered in two separate binding patches: one is formed by the N-terminal of helix1, the C-terminal of helix4 and the loop between helix4 and helix5, and a second patch is located in helix6 (Fig.1B).

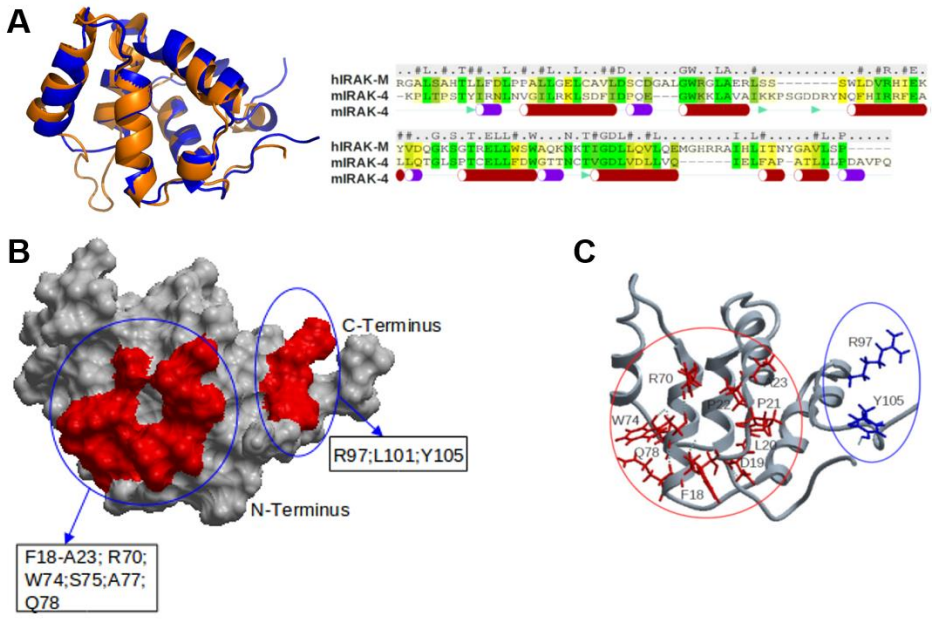


Figure 1. 3D structure model of the human IRAK-M death domain (DD). (A) The template 2A9I (Orange) was superimposed to the model (Blue). Sequence alignment of hIRAK-M-DD and mIRAK-4-DD. The sequence identity was 28.7%. Secondary structures such as alpha-helical and beta-strand of mIRAK-4-DD (2A9I) were denoted underneath the sequences (Red bar: alpha-helix: Purple bar: unstable helix; Green arrow: beta-strand). (B) Interactive surface prediction of hIRAK-M-DD. Space filling model with predicted interactive residues in red. (C) The

residues which were mutated in this study are shown with side chain and residue number in the back bone model and organized in patch number 1 (red) and patch 2 (blue).

Figure 2. Analysis of 100 ns molecular dynamics (MD) simulation of hIRAK-M-DD model. (A) The total energy of DD during MD simulation. (B) RMS deviations of Ca (blue), backbone (red) and heavy chain (yellow). (C) The number of hydrogen bonds in DD during the simulation. (D) The backbone RMS fluctuation of each residue of DD was calculated, on the X-axis the residue number from R8 to P111. The predicted residues contributing to protein interaction were less flexible (RMSF ~ 2 Å) and are shown in the red boxes.

Mutation/deletion of IRAK-M-DD

Based on our IRAK-M-DD model we probed potential interactive residues via structurally conservative mutagenesis of full-length human IRAK-M (Table 1). The selected residues for mutation are surface exposed and do not form intra-molecular contacts. Mutated residues are depicted with side chains in Fig.1C. In the first binding patch formed by helix1/helix4+loop we constructed a F18A mutant, generated a combined D19N-L20A-P21A mutant and a combined

P22A-A23S mutant, and R70Q, W74A and Q78A single mutant molecules. We verified that all mutants were expressed in 293T cells at a comparable level by evaluation of their expression levels with an antibody against the C-terminus of IRAK-M (Fig.3A). We next determined the capacity of human IRAK-M variants to activate NF-κB when overexpressed in 293T cells as described (1). As shown in Fig.3B human WT IRAK-M induces NF-κB in 293T cells, this function is fully dependent on the death domain since complete deletion of the death domain reduced this capacity to control level. Mutation of only F18 or D19-P21 only moderately affected the capacity of IRAK-M to induce NF-κB activity. Mutation of P22-A23 and Q78 caused a marked reduction in NF-κB. Single mutation of W74 completely abolished the capacity of IRAK-M to induce NF-κB in agreement with the study of Zhou *et al* (9) of murine IRAK-M. Combined mutation of F18/D19-P21 did not result in a further reduction of NF-κB compared to the single mutants (Fig.3B). Notably, the combined mutation of D19-P21/P22-A23 resulted in a mutant that regained its full NF-κB activating capacity as compared to the P22-A23 mutant, indicating that the D19-A23 stretch may harbor a negative control element. Combined mutation of F18/Q78, adjacent residues in the 3D structure, resulted in a mutant that essentially lost its NF-κB activating capacity (Fig.3B). Thus it seems that W74 and F18/Q78 interactions are pivotal in the NF-κB activating capacity of IRAK-M in 293T cells and that the D19-A23 stretch has a regulatory role in this.

Mutations in the predicted second interactive patch formed in helix6 by R97 and Y105, revealed that also R97 is of major importance in the NF-κB activating activity of IRAK-M while single mutation at Y105

results in only minor reduction (Fig.3B). However, combined R97/Y105 mutation reduced NF-κB to the level of the death domain deletion mutant which identifies these residues as a second crucial binding site for the NF-κB activating activity of IRAK-M (Fig.3B).

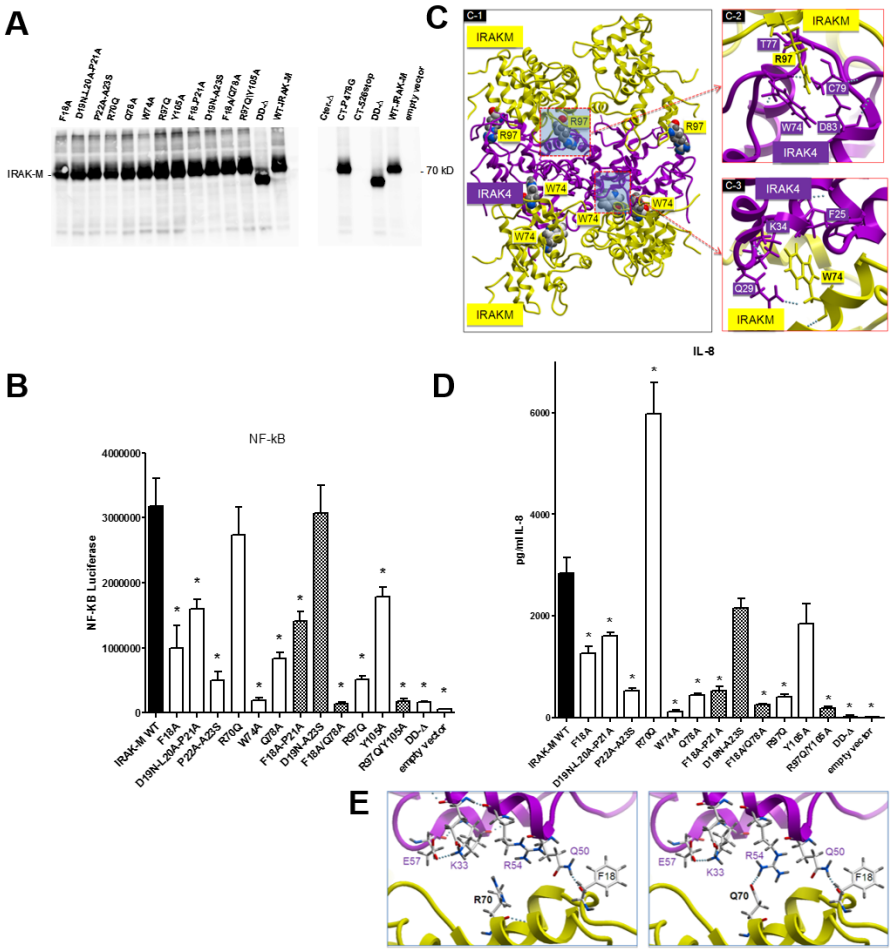


Figure 3. Expression and functioning of human IRAK-M-DD mutants in 293T cells. (A) Transient expression of IRAK-M and mutants by transfection in 293T cells by Western blotting performed on cell lysates with an antibody directed to the C-terminal of IRAK-M. (B) Effect of IRAK-M-DD mutations on NF-κB activation by overexpression in 293T cells. N=4, mean±SEM. * indicates difference with WT IRAK-M P<0.05. Shaded bars depict results of IRAK-M molecules with combinations of mutated residues/stretches within the same patch. (C) Potential

sandwich of one IRAK-4-DD tetramer by 2 IRAK-M-DD tetramers by W74 and R97 mediated interactions. The respective predicted IRAK-4 interaction points are shown in detail. (D) Effect of IRAK-M-DD mutations on IL-8 production by overexpression in 293T cells. N=4, mean±SEM. * indicates difference with WT IRAK-M P<0.05. Shaded bars depict results of IRAK-M molecules with combinations of mutated residues/stretches within the same patch. (E) Docking of mutant IRAK-M-DD predicts increased IRAK-4 interaction with the IRAK-M R70Q mutant through an extra hydrogen bond formed between Q70 in IRAK-M and R54 in IRAK-4.

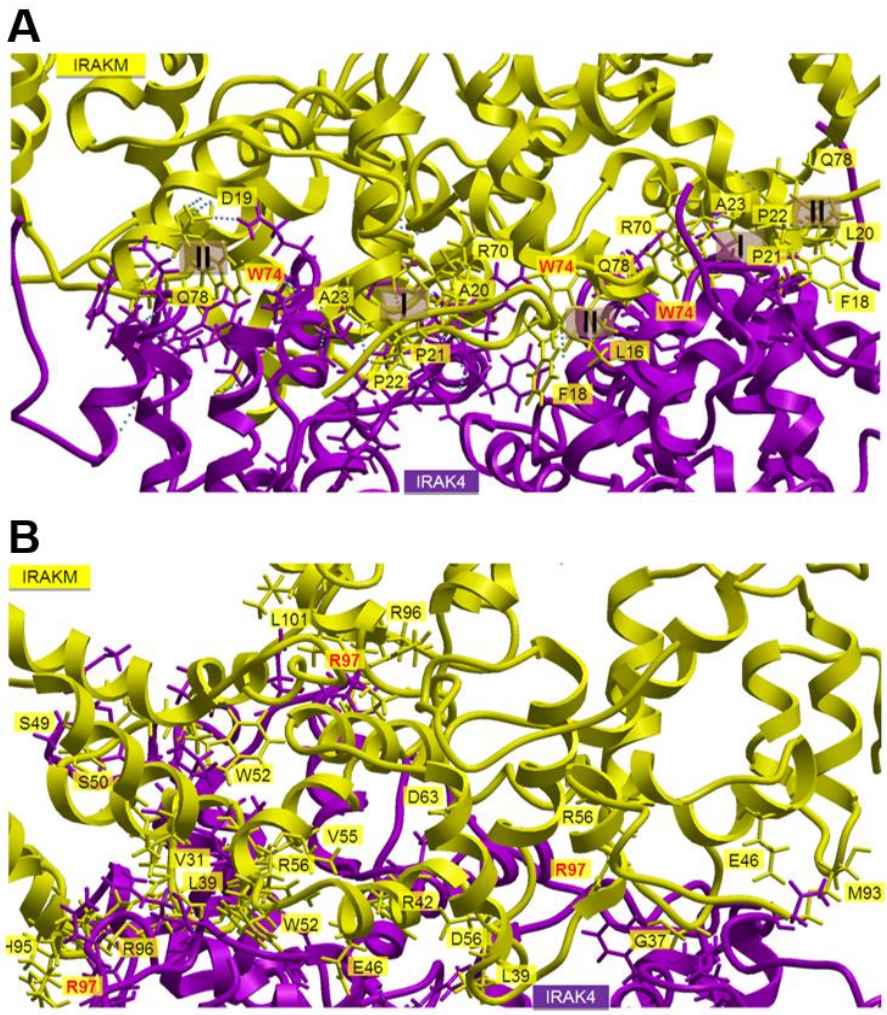


Figure 4. Interaction of IRAK-M and IRAK-4 tetramers. A. IRAK-M-DD interaction points predicted to be important for IRAK-M/IRAK-4 interaction by unbiased docking of the IRAK-M tetramer (Yellow) on the top surface of IRAK-4

tetramer (Purple). The residues involved in the interaction are shown with their side chains, of which the residues from IRAK-M are labeled with residue name and number. Two interaction types (I and II) were involved in the interaction and in the type I the residues involved are L20, P21, P22, A23, R70 while in the type II the residues involved are L16, F18, W74, S75, Q78. 63% of the 100 best docked modes maintained this binding mode. B. IRAK-M-DD interaction points predicted to be important for IRAK-M/IRAK-4 interaction by unbiased docking of the IRAK-M tetramer (Yellow) on the bottom surface of IRAK-4 tetramer (Purple). The residues involved in the interaction are shown with their side chains, of which the residues such as R97 involved for the interaction. 37% of the 100 best docked modes maintained this binding mode.

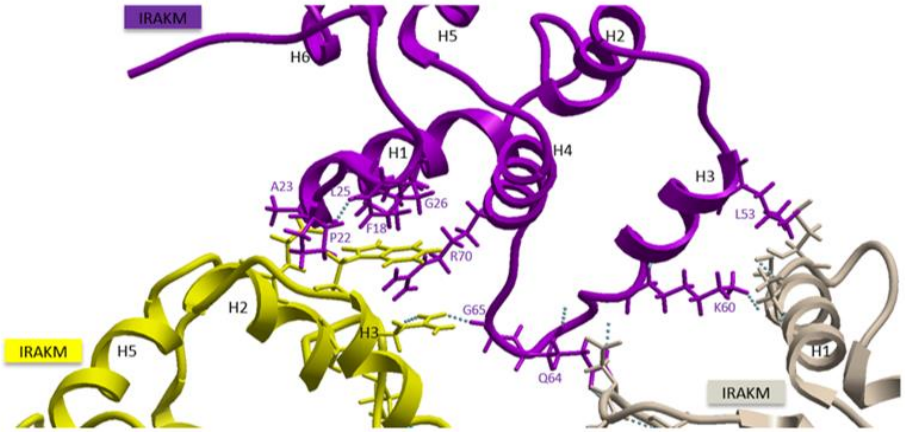


Figure 5. IRAK-M-DD interaction points predicted to be important for homomeric tetramer formation based on the crystallographic model of human MyD88/IRAK-4/IRAK-2 complex (3MOP). The binding sites of one IRAK-M momomer (purple) to two other IRAK-M monomers (yellow and grey) in the IRAK-M tetramer.

The observation that both W74 and R97 are pivotal for the IRAK-M activity towards NF-κB prompted us to model the interaction of our IRAK-M-DD structure with IRAK-4 (Fig.4) based on the earlier experimentally determined myddosome structure containing homomeric tetramers of IRAK-4 and IRAK-2 (3MOP.pdb (3)). In this structure, a tetramer of IRAK-4-DD's interacts with a tetramer of the DD's of MyD88 on one side, and a tetramer of IRAK-2-DD's on the

other side (3). In analogy to this constellation, and as suggested by the authors (3), the IRAK-M-DD's may also form homomeric tetramers as predicted based on its homology with IRAK-4 as well as IRAK-2 in the 3MOP structure.

From study of our model structure (Fig.1) we propose that IRAK-M-DD tetramers probably form by interaction of F18, P22, P23, L25, G26 and R70 of one IRAK-M-DD to L53, K60, Q64 and G65 of another DD (Fig.5). Unbiased *in silico* protein docking of the IRAK-M-DD tetramer onto the IRAK-4-DD tetramer side which interacts with IRAK-2 in 3MOP displays the W74 dependent interaction of IRAK-M with IRAK-4 (Fig.4A) in accordance with the reported W74 importance for IRAK-4 binding (9). However no interaction point is predicted for R97 in this type of interaction given that R97 is in fact located at the opposite side of the W74 interacting tetramer surface. Unbiased docking of the IRAK-M tetramer to the free side of the IRAK-2 tetramer in 3MOP also indicated the W74 side as interactive, without involvement of R97 (Fig.6A).

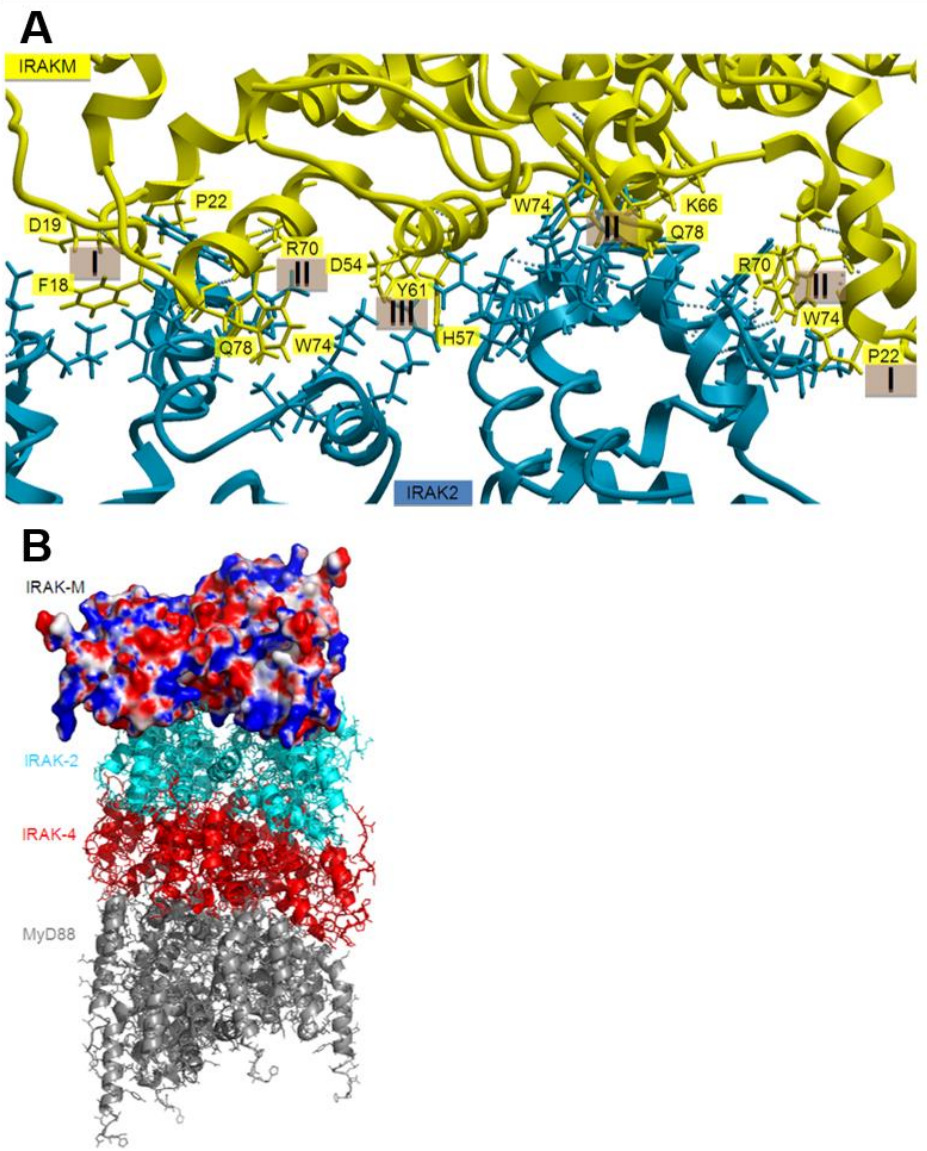


Figure 6. Interaction of IRAK-M and IRAK-2 tetramer. (A) IRAK-M-DD interaction points predicted to be important for IRAK-M/IRAK-2 interaction by unbiased docking of the IRAK-M tetramer on the free side of the IRAK-2 tetramer in the myddosome (3MOP) (B). In A part of the composite binding site of the IRAK-2 tetramer is shown with three IRAK-M molecules. The contact residues are shown with their side chain. The residues involved in the interaction from IRAKM are labeled. Three interaction types (I, II and III) were involved in the interaction.

However when the R97 exposing side of the IRAK-M tetramer was

docked unbiased at the top side of the IRAK-4 tetramer (Fig.4B) there was significant binding and good affinity prediction with R97 interacting with IRAK-4 residues W74, T77, C79 and D83 (Fig.3C). When IRAK-M tetramers were docked on IRAK-4 tetramers completely without any restriction, we observed two different overall docking poses: 63% of the binding events were with the W74 exposing site (see Fig.4A) and 37% with the R97 site (see Fig.4B). Furthermore, this suggests that IRAK-4 tetramers may actually form a complex with two IRAK-M tetramers on either side as depicted in Fig.3C. This sandwich hypothesis appears as most likely working model consistent with the notion that both W74 and R97 are crucial for the NF-κB activating activity of IRAK-M.

Overexpression of IRAK-M in 293T cells induced IL-8 secretion which involves NF-κB activation and postranscriptional regulation. The IL-8 production induced by the different DD-mutants (Fig.3D) occurred analogously to the NF-κB activity (Fig.3B), however with a few exceptions. The moderate effects of single mutation of F18 or D19-A21 on NF-κB were not associated with a further decrease of NF-κB in case of combined F18/D19-A21 mutation, while IL-8 production by the F18/D19-A21 mutant was markedly lower as compared to the single mutants. Furthermore, the R70Q mutant displayed hyperactivity with regard to IL-8 production compared to WT IRAK-M. Structure analysis of the DD-model indicated that the R70Q variant may actually form an extra hydrogen bond with R54 in IRAK-4 (Fig.3E). R70 is also involved in the interaction with IRAK-2 but different from IRAK-4 the Q70 mutant has no capabilities to form an extra hydrogen bond with IRAK-2, in contrast the Q70 mutant is

predicted to loose part of its interaction with IRAK-2. These results indicate a regulatory role of F18/D19-A21 and R70 in the way IRAK-M influences transcription/translation downstream or beside NF-κB.

Mutation/deletion of the C-terminal domain of IRAK-M

IRAK-M dependent NF-κB activation proceeds in IRAK-1/IRAK2 double deficient cells through a unique MEKK3 pathway (9). Taking this into account it could be hypothesized that IRAK-4/IRAK-M complexes recruit MEKK3 and activate TRAF6 in a specific manner that is dependent on a putative TRAF6 binding motif in the IRAK-M C-terminal domain (CTD) that extends from the inactive kinase domain. Because no experimental data existed yet on the functional involvement of the CTD (amino acid S445-E596) of IRAK-M, we generated an IRAK-M variant truncated at position S445 that lacks the entire CTD (CTD-Δ). Furthermore we mutated the $P^{478}VEDDE^{483}$ TRAF6 binding motif (10) in the CTD by introduction of a P478G substitution and generated a mutant that lacks the C-terminal part of the CTD by truncation at position K526. The P478G mutant was expressed similar as WT-IRAK-M, and consistently the CTD deletion and truncation mutant were not recognized by antibodies directed against the very C-terminus of IRAK-M (Fig.3A). A polyclonal antibody raised against full-length IRAK-M showed the expression of these CTD mutants (Fig.7A). Deletion of the complete CTD resulted in a major reduction of the NF-κB activating capacity of IRAK-M when overexpressed in 293T cells, while mutation P478G in the TRAF6 binding motif led to only a subtle if any reduction in NF-κB (Fig.7B). Truncation of the CTD at position K526 did not affect NF-κB activity

which indicates location of an important motif between position S445 and K526. These experiments indicate a major involvement of the CTD in the NF-κB activating capacity of IRAK-M, which seems independent of the TRAF6 binding motif at P478.

The importantly reduced capacity of the IRAK-M CTD-Δ mutant to activate NF-κB was associated with an almost complete loss of IL-8 production (Fig.7C). Interestingly, the TRAF6 binding motif at P478 appeared essential for IL-8 production by IRAK-M while NF-κB was hardly affected by the P478G substitution (Fig.7B-C). For comparison the DD-mutants D19-A21 and P22A-A23S mutant were studied simultaneously. These mutants showed similarly decreased NF-κB levels as respectively the P478G mutant and CTD-Δ mutant, however the CTD mutants displayed a much larger effect on IL-8 production compared to these DD-mutants. The minor effect of the P478G mutation on NF-κB and the large effect on IL-8 production indicate important other and distinct functions of the TRAF6 binding site in IRAK-M besides the function of the CTD in NF-κB activation.

IRAK-M is a MEKK3 pathway activator (9) and, consistent with the notion that MEKK3 activates NF-κB and ERK1/2 specifically in 293T cells (22), we also observed ERK1/2 activation upon overexpression of IRAK-M in these cells. We determined ERK1/2 phosphorylation for cells transfected with the IRAK-M-DD and CTD mutants as an indication of MEK activation (Fig.7D). The DD-mutants that were virtually devoid of NF-kB activating activity also failed to activate ERK1/2 (Fig.7D). However the DD mutants that were relatively hyperactive towards NF-κB (D19-A23) and IL-8 (R70Q) were

hampered in their capacity to activate ERK1/2 (Fig.7D). Also the TRAF6 binding motif P478G mutant showed a marked reduction in phospho-ERK1/2 generation. Complete deletion of the CTD resulted in a major decrease of ERK1/2 activation while CTD truncation at position K526 was without effect. These results indicate a special involvement of R70, D19-A23 and P478 in the MEK activating activity of IRAK-M towards ERK1/2.

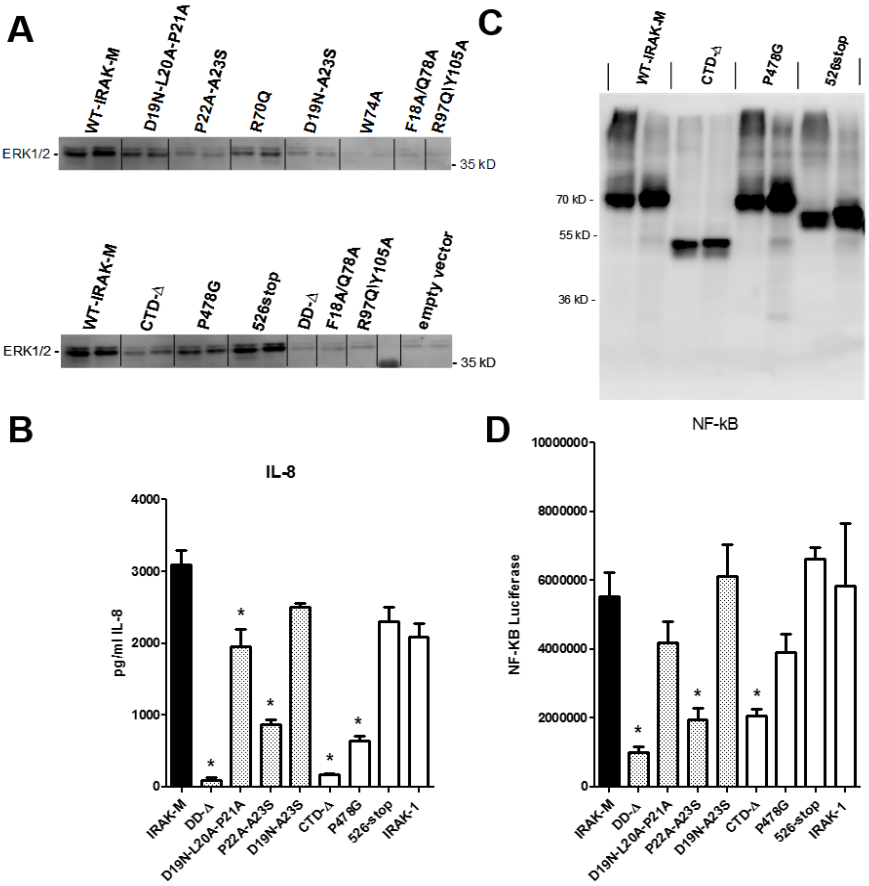


Figure 7. Expression and functioning of IRAK-M C-terminal tail mutants and IRAK-M-DD mutants in 293T cells. (A) Transient expression of IRAK-M and C-

terminal tail mutants by transfection in 293T cells by Western blotting performed on cell lysates with an antibody directed to full length IRAK-M. (B) Effect of IRAK-M C-terminal truncation and mutation compared to IRAK-M-DD mutation on NF-κB activation by overexpression in 293T cells. N=4, mean±SEM. * indicates difference with WT IRAK-M P<0.05. (C) Effect of IRAK-M C-terminal truncation and mutation compared to IRAK-M-DD mutation on IL-8 expression by overexpression in 293T cells. N=4, mean±SEM. * indicates difference with WT IRAK-M P<0.05. (D) Effect of DD and C-terminal mutation on the ERK1/2 activating activity of IRAK-M by overexpression in 293T cells. Westernblotting on cell lysates with anti-pERK1/2.

Potency of IRAK-M-DD mutants to inhibit TLR signaling in macrophages

IRAK-M is expressed in monocytes/macrophages and in lung epithelial cells and IRAK-M is an important factor to down-regulate the host defense in bacterial pneumonia models (1, 14-16, 19). In order to study the potential pro- and anti-inflammatory effects of human WT IRAK-M and the DD mutants on TLR mediated cytokine/chemokine release we stably introduced them in human monocytic cells (THP-1) and in human bronchial lung epithelial cells (H292). Coding sequences of WT and mutant IRAK-M were stably introduced by a lentiviral system *in trans* with eGFP that enabled FACS-sorting of eGFP positive cells of the same high intensity to obtain identical and homogeneous transcription of WT and mutant transgenes. Most mutants showed comparable or somewhat higher protein levels than WT IRAK-M upon stable expression in THP-1 and H292 cells (Fig.8A and 9A), however the hyperactive mutant R70Q displayed lower expression, and the D19N-A23S mutant displayed markedly lower steady state protein levels. Identical levels of IRAK-M mRNA were observed for the R70Q and D19N-A23S mutants as WT

(Fig.10), and since all mutants showed similar protein expression upon transient expression in 293T cells (Fig.3A) it appears that the R70Q and D19N-A23S mutants are prone to increased protein turnover in a more proficient cell type.

Inhibition studies with proteasome inhibitor MG-132 or an IRAK-1/4 inhibitor did not normalize expression of these mutants. Thus the mechanism underlying the high turnover and/or low protein expression level of these mutants remains elusive, although it seems to correlate with their relative observed hyperactivity.

Overexpression of IRAK-M in macrophages significantly inhibited TLR2 and TLR4 induced TNF and IL-6 production in a death domain dependent manner (Fig.8B). Mutations in interactive patch 1 formed by residues F18, D19-P21, P22-P23, W74 and Q78 resulted in partially or completely restored cytokine production in macrophages (Fig.8B). Mutations in the predicted interactive patch 2 in IRAK-M formed by R97 and Y105 also released the restriction on TLR2 and 4 mediated IL-6 production, while R97 and Y105 were not, or less implicated in the inhibitory action of IRAK-M on TNF production. Although the steady state level of the hyperactive D19-A23 mutant is low in macrophages (Fig.8A) this mutant still reduces LPS mediated TNF and IL-6 production (Fig.8B). These results indicate that IRAK-M actively inhibits cytokine production in macrophages with W74 and R97 as key residues.

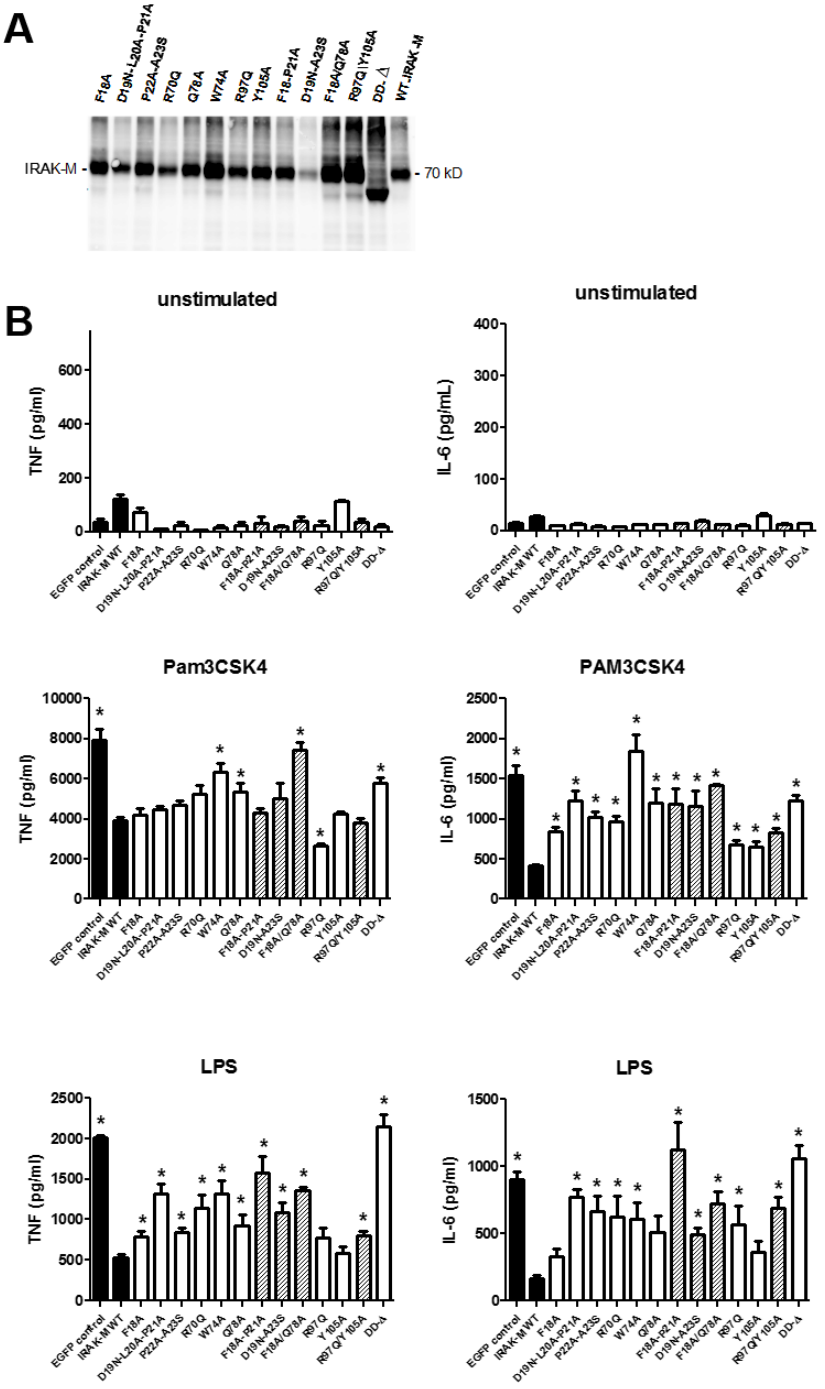


Figure 8. Effect of human IRAK-M and death domain mutants in

macrophages. For stable overexpression with IRAK-M and DD-mutants in monocytes/macrophages the human monocytic cell line THP-1 was transduced and FACS-sorted on EGFP positivity which is expressed in trans with WT IRAK-M and DD-mutants in the transgene by an IRES. THP-1 cells were matured to macrophage before stimulation as described in Methods. (A) IRAK-M expression was evaluated by western blotting. (B) The effect of stable IRAK-M expression and IRAK-M mutants was determined on TLR2 (PAM3CSK4) and TLR4 (LPS) mediated TNF and IL-6 production (B) after stimulation for 6 hour. Shaded bars depict results of IRAK-M molecules with combinations of mutated residues/stretches within the same patch. N=4, mean±SEM. * indicates difference with WT IRAK-M $P<0.05$.

Potency of IRAK-M-DD mutants to inhibit IL-1 and TLR5 signaling in lung epithelial cells

IRAK-M has been reported to be present in lung epithelial cells (19). Lung epithelial cells are relatively unresponsive to TLR2 and TLR4 agonist, but react potently to MyD88/IRAK dependent IL-1 receptor and TLR5 stimulation ((23), and own observation in H292 cells).

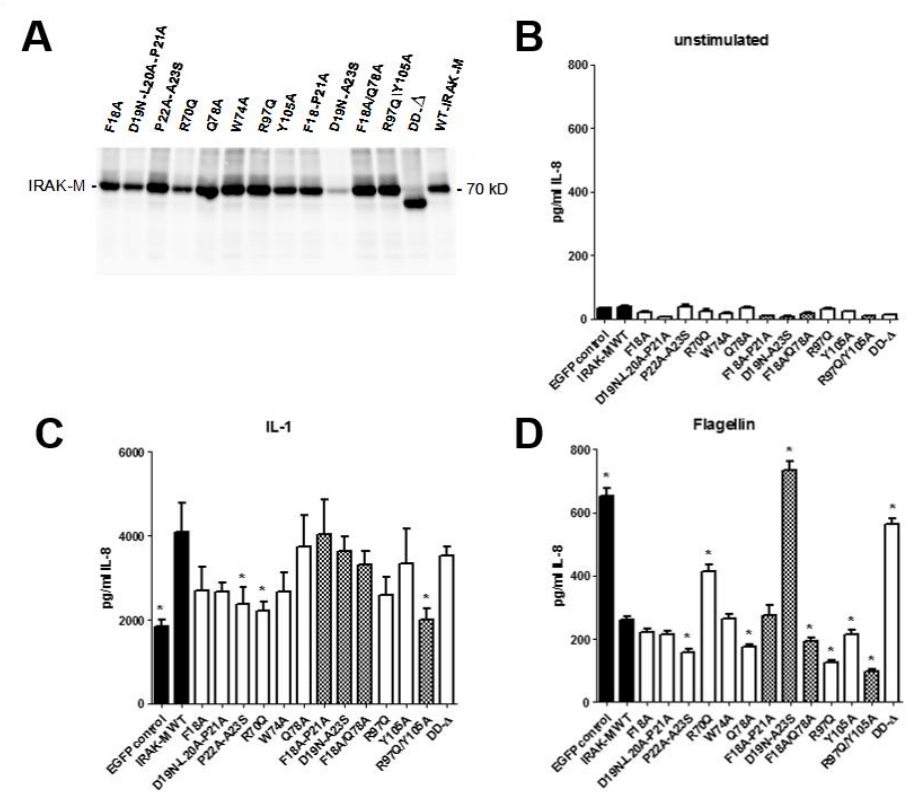


Figure 9. Effect of human IRAK-M and death domain mutants in lung epithelial cells. For stable overexpression with IRAK-M and DD-mutants the human bronchial lung epithelial cell line H292 was transduced and FACS-sorted on EGFP positivity expressed in trans with WT IRAK-M and DD-mutants in the transgene by an IRES as described in Methods section. (A) IRAK-M expression was evaluated by western blotting. (B) The effect of stable IRAK-M expression and IRAK-M mutants was determined on IL-1β and Flagellin (TLR5) mediated IL-8 expression in the supernatant after 6 hour stimulation. Shaded bars depict results of IRAK-M molecules with combinations of mutated residues/stretches within the same patch. N=8, mean±SEM. * indicates difference with WT IRAK-M $P<0.05$.

Most prominent inflammatory mediator produced by these cells is IL-8 and IL-8 mRNA is already relatively abundant in non-stimulated lung epithelium (24). Especially TLR5 mediated IL-8 production in epithelial cells is regulated via posttranscriptional mechanisms (25). TLR5 is activated by Flagellin derived from the flagella of certain Gram-negative bacteria such as *P. aeruginosa.* In this respect it is

noteworthy that IRAK-M$^{-/-}$ mice are protected from *P. aeruginosa* pneumonia under immunocompromised conditions caused by cecal ligation and puncture (14). TLR5 mediated IL-8 production by H292 lung epithelial cells was significantly inhibited by WT IRAK-M (Fig.9B).

The inhibition of TLR5 mediated IL-8 production by IRAK-M was completely death domain dependent since the DD-deletion mutant showed no effect on IL-8 production. Mutation of R70 partially restored IL-8 production and mutagenesis of the D19-A23 stretch resulted in complete restoration of chemokine production. Interestingly mutation of key residue W74 did not influence the inhibitory effect exerted by IRAK-M on TLR5 mediated IL-8 production, and mutation of R97 and Y105 even enforced this inhibitory capacity compared to WT IRAK-M (Fig.9B), an effect that was also observed for some patch 1 mutants. Remarkably, IL-1β stimulated IL-8 expression by H292 lung epithelial cells was enhanced by WT IRAK-M (Fig.9B), which is however consistent with the notion that IRAK-M may substitute for IRAK-1 in IL-1β signaling (1). Most DD-mutants did not show this stimulatory effect, interestingly however the DD-deletion mutant which lacks the entire DD also stimulated IL-1β driven IL-8 expression.

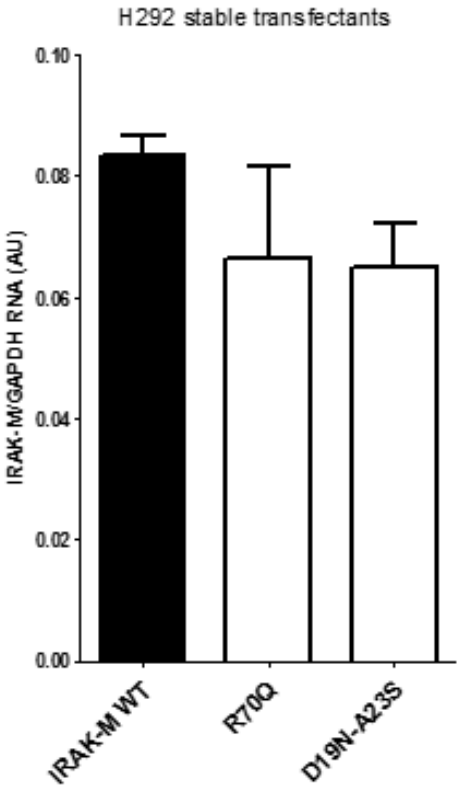


Figure 10. Stable expression of R70Q and D19N-A23S mutants in H292 cells as evaluated at the mRNA level by RT-qPCR. Mutants are transcribed equal to WT IRAK-M upon stable overexpression. The low observed protein levels upon stable expression in proficient cells is not due to lower transcription.

To evaluate whether the dampening of TLR5 mediated IL-8 production by IRAK-M was caused by increased expression of negative feedback inhibitors such as A20 as described (9) we determined A20, SHIP-1 and SOCS-3 expression in H292 cells that overexpressed WT or mutant IRAK-M proteins after 3 hours of stimulation. IRAK-M however decreased flagellin induced A20 and

SOCS-3 expression in H292 cells (Fig.11), and while IRAK-M upregulated basal SHIP-1 levels this function was retained by the DD-deletion mutant that did not inhibit IL-8 release. Thus it appears that IRAK-M regulates TLR5 mediated responses in lung epithelial cells by inhibition of the forward signaling to IL-8 induction, not by increase of negative feedback inhibitors. It should be noted however that IL-1β induced A20 expression was increased in the case of the W74A mutant compared to WT IRAK-M while A20 expression was reduced for the R97 and R97/Y105 mutants. Also flagellin stimulated A20 production was importantly higher for the W74 mutant compared to WT and the R97Q mutant. Thus it appears that A20 expression is increased when the R97 dependent interaction of IRAK-M is favored. These results indicate that residue R97 may be implicated in the IRAK-M dependent expression of negative regulators of inflammation such as A20.

The mechanism by which IRAK-M inhibits TLR5 mediated IL-8 responses in lung epithelial cells (Fig.9B) is apparently different from the mechanism by which IRAK-M inhibits TLR2 and TLR4 mediated TNF and IL-6 production by macrophages (Fig.8B) since different residues are involved in the inhibitory function as shown by these mutagenesis experiments. IRAK-2 dependent stimulation of translation is potentially crucial for TLR5 mediated IL-8 production by epithelial cells (25) which will be inhibited by IRAK-M (9), according to our protein docking experiments, in a R97 independent manner. Since the R97Q mutant will be less occupied with IRAK-4 it is not unlikely that more of this mutant will be available for IRAK-2 binding and consequently may exert a more pronounced effect on IL-8

production.

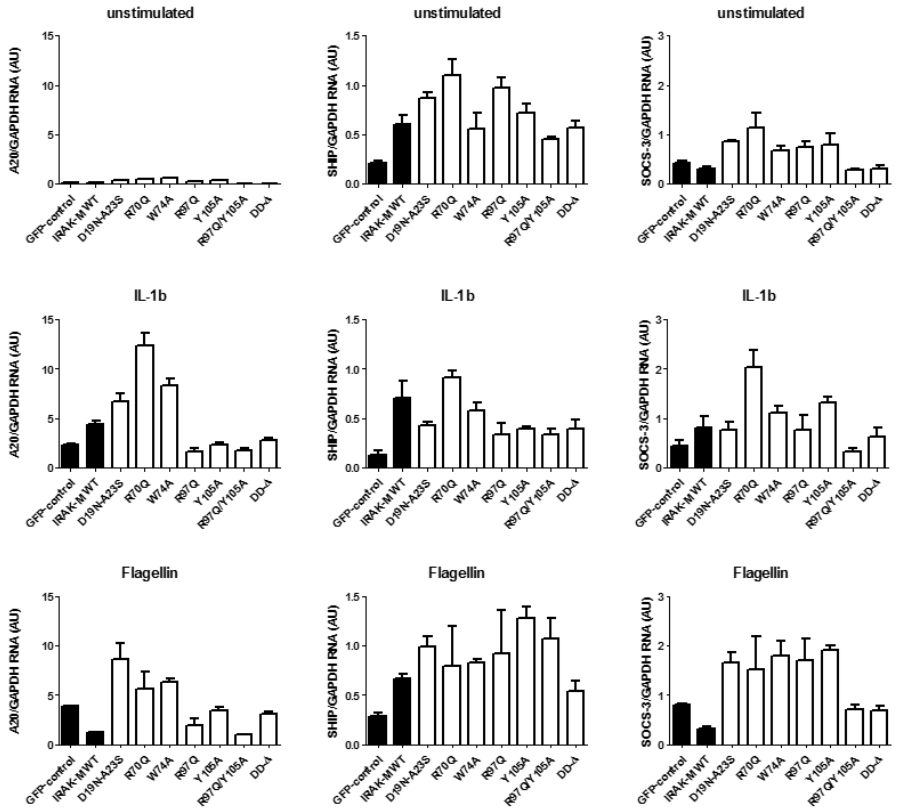


Figure 11. Effect of over expression of the different IRAK-M death domain mutants in H292 cells on the expression of other negative regulators. H292 cells were stimulated for 3 hours with IL-1b (2 ng/ml) or flagellin (25 ng/ml) for 3 hours and mRNA expression was evaluated by RT-qPCR as described under materials and methods.

Effect of IRAK-M mutants on TRAF6 expression

To evaluate potential effects of IRAK-M mutation on TRAF6 function we co-expressed IRAK-M with HA-tagged TRAF6 in 293T cells. IRAK-M co-expression increased TRAF6 protein levels (Fig.12). This

function was strongly enhanced when the IRAK-M death domain is deleted, but apparently independent of the deemed pivotal P478 residue in the TRAF6 binding consensus (Fig.12). While expression of the death domain deletion mutant itself does not lead to IL-8 production, it does so when co-expressed with low amounts of TRAF6 vector (Fig.12), indicating that this phenomenon may increase pro-inflammatory signaling. These results indicate positive effects of IRAK-M expression on TRAF6 stability that is independent of the DD. This influence of IRAK-M on TRAF-6 and the exaggeration of this effect by deletion of the DD can be in part explanatory to the observed increase in IL-1β stimulated IL-8 production in H292 cells by the DD-deletion mutant (Fig.9B). Potentially TRAF6 levels are a limiting factor for IL-8 production upon stimulation of IL-1β of H292 cells and TRAF6 protein levels may have been increased in these cells by the DD-deletion mutant. Together these results indicate that death domain-less IRAK-M is not an inert molecule but may actually stimulate pro-inflammatory reactions when TRAF6 is a limiting factor. This effect may be well of physiological relevance because the death domain of IRAK-M can be cleaved of at D135 by caspase-6 when monocytes/macrophages contact activated neutrophils (26).

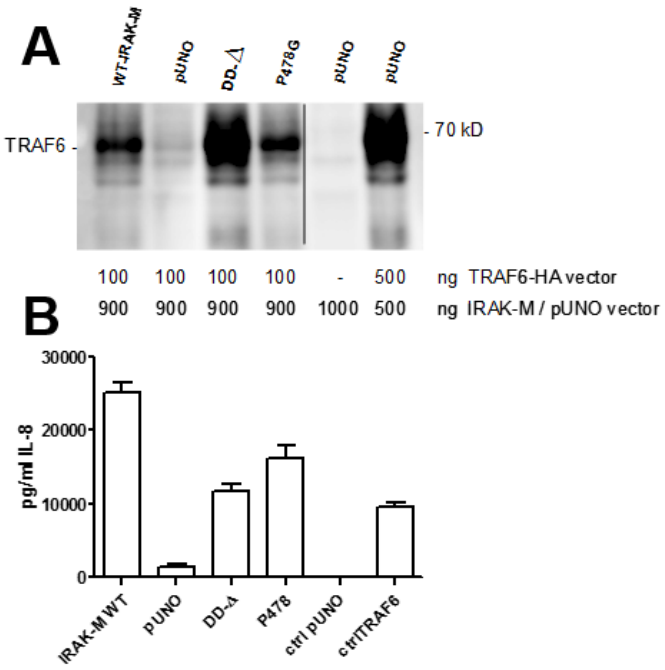


Figure 12. Coexpression of IRAK-M with HA tagged TRAF6 in 293T cells stabilized TRAF6 protein and deletion of the death domain enhanced this effect. A total of 1000 ng plasmid DNA was transfected, (-) is transfection with control pUNO plasmid only. Coexpression of the DD-deletion mutant with TRAF6 also resulted in IL-8 production with the TRAF-6 / DD-del cotransfection at a low TRAF6 vector concentration that generates no IL-8 when transfected alone.

Discussion

We generated a structure model for the death domain of human IRAK-M, a member of the IRAK protein family involved in IL-1/Toll-like receptor signaling. The model for the hIRAK-M-DD was analyzed by a consensus approach of several structural bioinformatics techniques in order to predict the most likely interaction areas that are involved in IRAK-M ligand binding and we identified 2 areas (Fig.1B) on the IRAK-M-DD that are primary protein-protein

interaction regions. The model was used to rationalize the targeted mutagenesis of IRAK-M in order to obtain novel structure function information. Mutations in either of the predicted interaction regions led to alteration of function of the mutated IRAK-M proteins when expressed in HEK293T cells, monocyte/macrophage cells (THP-1) and bronchial lung epithelial cells (H292). Residue W74 in patch 1 was found to be crucial for NF-κB activation while R97/Y105 were found essential for activity towards NF-κB provided by patch 2. Interestingly W74 and R97 are predictably located on opposite sides of the IRAK-M tetramer when IRAK-M forms homomeric tetramers via its DD in structural analogy to the experimentally determined quarternary structures of the DD of IRAK-4 and IRAK-2 in the myddosome (3). Because IRAK-M activity towards NF-κB is IRAK-4 dependent (7), we hypothesized that IRAK-M tetramers will interact with IRAK-4 in a W74 and R97 dependent manner. Protein docking experiments with a predicted IRAK-M-DD tetramer indicated that the W74 and R97 sides of an IRAK-M tetramer may bind respectively to the bottom and top side of an IRAK-4 tetramer (bottom and top as defined in Ref.3). Since both W74 and R97 are essential for the NF-κB activating activity upon overexpression in 293T cells it appears that it is actually the IRAK-M/IRAK-4/IRAK-M sandwich that generates the NF-κB activity in this system. Importantly, analysis of IRAK-M variants also showed that both W74 and R97/Y105 were involved in the IRAK-M mediated inhibition of TLR2 and TLR4 driven IL-6 production by human macrophages. Furthermore, IRAK-M residue W74 appeared essential in down-regulation of TNF, but R97 not. W74 and R97 were not implicated in the inhibition by IRAK-M of

TLR5 mediated IL-8 production by lung epithelial cells which could be consistent with the notion that TLR5 mediated IL-8 production is regulated postranscriptionally (25) and posttranscriptional regulation by IRAK-M is effectuated by its action on IRAK-2 (9). In contrast, the region located in between W74 and R97 was important for this function, specifically R70 and D19-A23. Interestingly increased A20 expression was not associated with the inhibitory effect of IRAK-M in lung epithelial cells as reported (9), however A20 expression appeared to be associated with the R97 type of interaction of IRAK-M.

Further inspection of the functionality of the IRAK-M mutants involved ERK1/2 phosphorylation which is driven by the MEK pathways. Human IRAK-M activates ERK1/2 upon overexpression in 293T cells, and all DD-mutants that were hampered in their NF-κB activating activity and IRAK-4 binding were also hampered in ERK1/2 activation. However R70 and D19-A23 seem to be specifically involved in ERK1/2 activation since mutation of these residues showed equal NF-κB activation but greatly reduced phospho-ERK1/2. Also the TRAF6 binding site (P478) in the CTD seems specifically involved in ERK1/2 phosphorylation and not in NF-κB. Thus it appears that the area in between the pivotal residues for NF-κB activation, namely R70 and D19-A23 are specifically involved in mediating the P478 dependent ERK1/2 signaling. This potentially requires specific binding or orientation of MEK kinase by the R70/D19-A23 area of IRAK-M.

From the overexpression experiment in 293T cells it could be

concluded that residues D19-A21 and P22-A23 are involved in both positive and negative effects on signaling events. This is consistent with our finding that the D19-A21 as well as the P22-A23 mutant proteins lack capacity to inhibit cytokine expression in macrophages (Fig.8B).

Interestingly, the relative NF-κB hyperactivity of the D19-A23 IRAK-M mutant (Fig.3B and 7B) is associated with lower protein levels in proficient cells (Fig.8A and 9A). For IRAK-1 it has been suggested that an intramolecular interaction of the death domain with the end of its CTD keeps IRAK-1 in a silent mode before phosphorylation events trigger its activation (27). IRAK-1 that lacks the end of its CTD is easily activated and instable (27). In the context of IRAK-M one could hypothesize a role of the D19-A23 stretch in the stability of IRAK-M either by interaction with its own CTD or specific interaction/repelling with/of other IRAK or associated molecules. The lower observed steady state protein levels of the IRAK-M mutant with the modified D19-A23 stretch (Fig.8A and 9A) would be consistent with faster turnover through increased kinetics of activation and subsequent degradation events. In this respect, blotting experiments with the IRAK-M K526stop mutant that lacks the C-terminal epitope recognized by the used C-terminal anti-IRAK-M antibody, showed clearly that both the low and high molecular bands observed upon expression of IRAK-M are specific IRAK-M derived products. These products appear to display the continuous modification and degradation of WT IRAK-M as well as the mutants (Fig.3A, 8A, 9A).

We compared the homologues residues involved in the action of

IRAK-M on the structures of other IRAK's (Fig.13). It appeared that the stretch F18-A23 in IRAK-M, which is involved in NF-κB and ERK1/2 activating activity as well as inhibitory activity and protein levels (Fig.3,7,8,9), displays the least homology with the other IRAK's (Fig.13). Major difference with IRAK-2 is that A23 in IRAK-M is a tryptophan in the homologous residue in IRAK-2 (W11) which provides a bulky mass to the exterior of binding patch 1 which is lacking in IRAK-M (Fig.13). Furthermore the negative charge provided here in IRAK-M by D19 is lacking in IRAK-2 (Q7). In IRAK-4 this negative charge is also lacking at the homologous position (C13), instead a positive charge is provided in IRAK-4 in this area by R12 in the homologous position of the aromatic phenylalanine (F18) in IRAK-M (compare Fig.13 B-1 and B-3). Undoubtedly these differences will contribute to the divergent functions of IRAK-M and the other IRAK's. The residues in the D19-A23 stretch are predicted IRAK-M interaction points with IRAK-M itself (A23, P22), with IRAK-4 (L20, P21, P22, A23), and IRAK-2 (D19, P22) (Fig. 4-6).

A naturally occurring P22L mutant is reported to be associated with early onset asthma (19). Residue P22 is predicted to be involved in IRAK-2/4 and M interactions as mentioned above. The enhanced propensity of individuals with the P22L mutation to develop asthma is in line with the observed decreased capacity of IRAK-M with mutations in this region (P22A-A23S mutant) to downregulate IL-6 expression in macrophages upon TLR2/4 stimulation (Fig.8B).

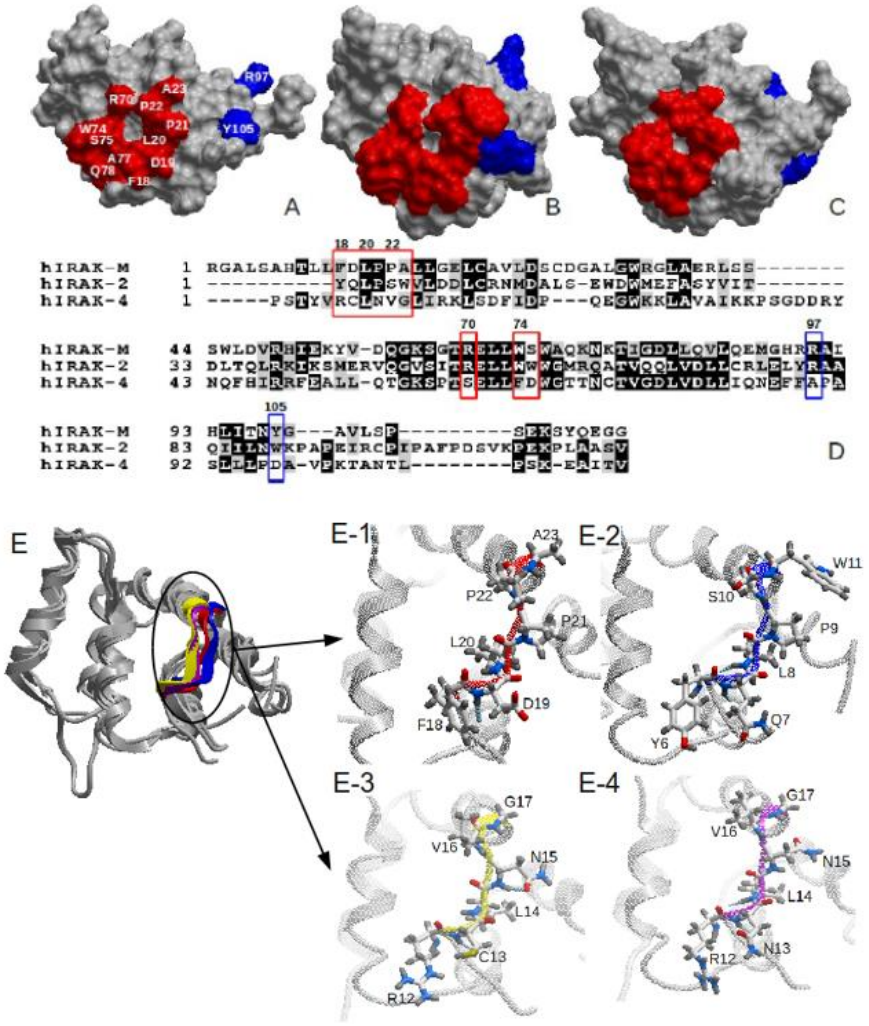


Figure 13. Comparison of the hIRAK-M Death Domain interactive surface with IRAK-2 and IRAK-4 DD's. Upper panel: Comparison of the defined interactive patches 1 (red) and 2 (blue) of the Death Domain of IRAK-M (A) to their homologous residues on human IRAK-2 (B) and human IRAK-4 (C). Death Domains that were placed in a similar orientation by 3D structural overlay. In D the multipe sequence alignments. The residues involved in patch 1 and patch 2 are shown in red and blue boxes, respectively. Superposition of IRAK members with the colored F19-A23 stretch (E). F18-A23 stretch and homologous residues shown in detail (E1-4). E-1:hIRAK-M, E-2:hIRAK-2, E-3:hIRAK-4, E-4:mIRAK-4.

Mutation of predicted interactive residues of the second binding site on the death domain of IRAK-M revealed that R97 plays a crucial role in this region and an additional role for Y105. R97 is conserved in IRAK-2, not in IRAK-4. Y105 appears to be unique for IRAK-M, and may be involved in the specific functionality of IRAK-M (Fig.13). A low frequent naturally occurring R97Q mutant has been reported (rs185025028). Here we show that the R97 residue is dominantly involved in NF-κB, ERK1/2 and inhibition of TLR elicited cytokines by macrophages. It will be interesting to see whether this genotype is associated with altered disease phenotype or outcome.

Zhou et al (9) showed that W74 is involved in IRAK-4 binding and NF-κB activating activity of IRAK-M when macrophages are stimulated with viral type TLR7 agonist. Here we show that the capacity of IRAK-M to reduce TLR2 and TLR4 mediated cytokine release depends on both W74 and R97. To our knowledge this is the first demonstration that W74 and R97 are involved in the inhibitory action of IRAK-M.

In vivo IRAK-M may be destroyed as an inhibitor by cleavage through caspase-6 at residue D135 which causes the complete removal of the death domain (26). Indeed, here we found that death domain-less IRAK-M no longer functions as an inhibitor. In contrast, we observed that IL-1 induced IL-8 production was significantly stimulated by death domain-less IRAK-M. In line we noticed that IRAK-M stabilizes TRAF6 when co-expressed and that this effect is enhanced in the absence of the death domain (Fig.12). Thus while removal of the death domain by cleavage at the residue D135 by caspase-6 may

inactivate certain activities of IRAK-M, it may at the same time stimulate other processes by increasing the TRAF6 concentration. That IRAK-M without the DD still displays some activity is consistent with our finding that the C-terminal domain of IRAK-M is fully functional and required for NF-κB and ERK1/2 activation (Fig.7).

It should be mentioned that we modeled the interactions of IRAK-M with the other IRAK's as if IRAK-M forms homotetramers and interacts as such with homotetramers of IRAK-4 and IRAK-2. This is based on the homotetramer formation observed for the isolated death domains of MyD88/IRAK-4/IRAK-2 when co-crystallized (3). Full-length proteins may however interact differently which is already exemplified by the importance of the intermediate domains immediately extending from the DD of MyD88 and IRAK-1 (28, 29). The mechanism by which IRAK-M exerts its special inhibitory actions will also depend on specific binding events and the differential activation and inhibition of TAK, TAB, MEKK and TRAF6 compared to the other IRAK proteins as outlined by Zhou et al (9) and again schematically represented in Fig.14 with our new findings incorporated. Elucidation of the binding sites important for IRAK-4/IRAK-M/MEKK3 interaction and the involvement of TRAF6 in activity of this complex will be key to understand how these mediators come to the specific inhibitory effect of IRAK-M.

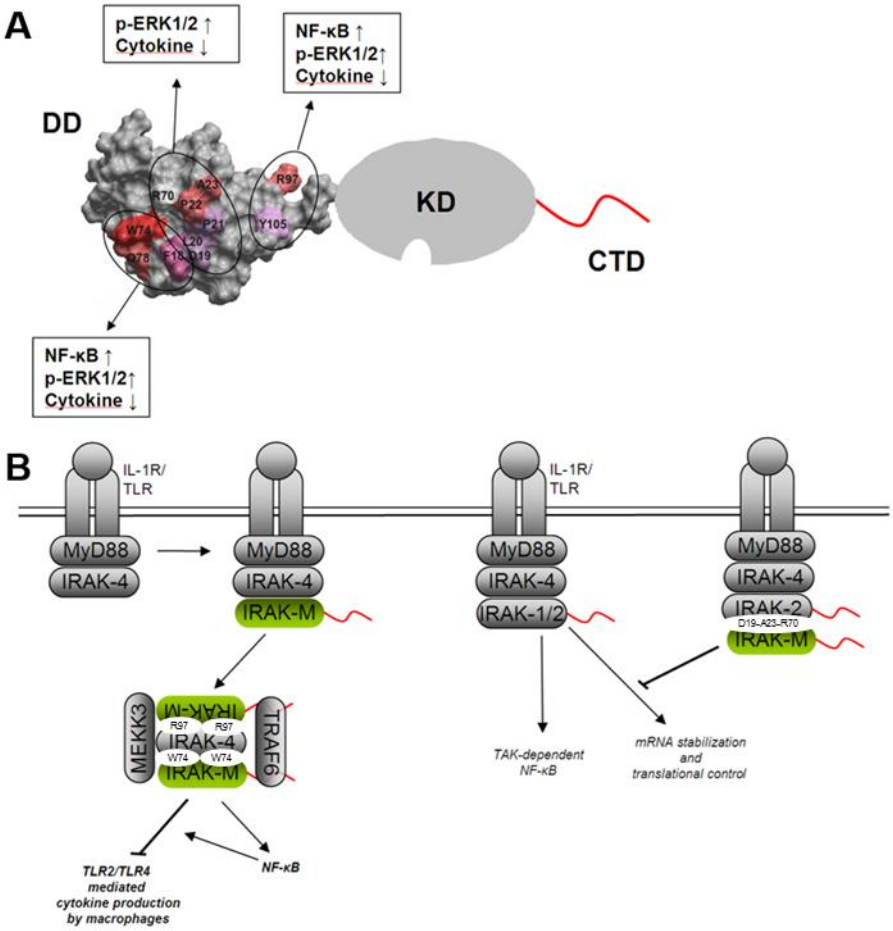


Figure 14. Working mechanism of IRAK-M as inhibitor based on specific residues of its death domain and C-terminal domain.

In conclusion, the present study elucidates a large part of the structure function relationships of the DD of IRAK-M. Our findings may guide targeting of the DD of human IRAK-M in efforts to generate treatment strategies to prevent bacterial pneumonia in immunocompromised patients.

Methods

Homology modeling and interactive surface prediction

Template Selection and Alignment

The primary sequence of human IRAK-M consists of 596 amino acids. In order to determine the domain boundaries of the death domain and in order to include the complete domain region, we applied multiple sequence analysis of IRAK-M from different species. The death domain could be modeled from C5 to G119 and potential templates for the IRAK-M death domain were selected through PSI-BLAST of the protein data bank (PDB) (30). PSI-BLAST (PSSM: 0.005) results showed that there were 12 coordinate files available, with sequence identities between 24% and 30% with the IRAK-M death domain sequence. Based on sequence alignment score and structure quality of potential template structures (resolution and R-factor) the final template was selected. We selected 2A9I (DD of mouse IRAK-4, Ref.8) as a template, based on its sequence identity (28.7%). Moreover, 2A9I describes the crystal structure of the death domain of IRAK-4 with a high resolution at 1.7 Å which was optimized by PDB_REDO server with a structural quality of 0.619. The alignment between template and the IRAK-M sequence was performed by ICM-Pro (MolSoft) (31) with default scoring parameters and refined based on secondary structure prediction, amino acid features and the 3D structure of the template (Fig. 1A and Table 2).

Backbone Generation and Loop Modeling

For the initial construction of the models we employed the ICM Pro molecular modeling package Molsoft (31) as well as LOOPY (32). ICM-Pro was used to generate the backbone coordinates for the models, from the refined sequence alignment between template and query structure, except in several variable (loop) regions, for which we used LOOPY. If the template and target had identical residues also the side chain coordinates could be included. Side chains were minimized by steepest descent and simulated annealing minimization with a fixed backbone conformation and next optimized without any restraint using the ICM Pro package.

MD simulation and Quality Check

The optimized model structure was further refined using a short MD simulation in explicit water (density 0.997, pH 7.0) for 500 picoseconds, employing the Yasara-Whatif twin package. The YAMBER3 force field was used, periodic boundaries and long range Coulomb interactions were included with a cutoff of 7.86Å. Every 25 ps, a simulation snapshot was saved, and in total, 20 snapshots were produced. Every snapshot and template were submitted for online structure quality check at http://nihserver.mbi.ucla.edu/SAVES_3/, using PROCHECK, WHATIF, VERIFY-3D, ERRAT and PROVE. The snapshot with the best total energy was selected as the starting point for a 100 ns MD simulation, with the AMBER03 force field periodic boundaries and long range Coulomb interactions were included with a cutoff of 7.86 Å in explicit water (Density 0.997, pH 7.0, NaCl 0.9%). The resulting energies, RMSD, residues' flexibility, hydrogen bonds and Ramachandran plots from the MD simulation were

calculated.

Evaluation of the structural quality (Table 3) of the final death domain model indicated that 92.1% of the residues were located in most favored zones and the remaining 7.9% were present in allowed regions as analyzed by inspection of a Ramachandran plot. Conformation Z-scores of both the model and the template were low, though the model's side chain planarity and inter atomic distances were good. The structure's non-bonding interactions were qualified as good in both the model and the template. The residues' environment in the model is not better than that in template, but both of them are qualified as reasonable. A 100 ns MD simulation was performed to analyze the structural stability, amino acid fluctuation, and potential energy changes of the IRAK-M death domain (Fig.2). During the MD simulation, the total energy (combined by bond, angle, planarity, coulomb, VdW) was stable while the conformation of the model changed during the simulation until 10 ns and then stabilized. The fluctuation of each residue during the MD simulation was calculated, and we observed that the loop between helix3 and helix4 was flexible. The number of hydrogen bonds, known to be important for protein stability and function (33), was generally consistent in the model during the simulation. A107 and N104 form hydrogen bonds with each other and this way connect helix 6 with the C-terminal loop.

Prediction of protein-protein contacts

A consensus approach was used to predict protein-protein interaction

for the model structure generated by application of several structural bioinformatics methods. We employed: Optimal Docking Area (ODA) (34), Cons-PPISP (35) and PPI-Pred (36). For each of these methods, a prediction score was obtained for all residues in the death domain of IRAK-M and a consensus was generated from the different methods applied, taking into account the accessibility of the residue. Several important residues for the interaction between the IRAK-4 death domain and the MyD88 death domain have been predicted on basis of an earlier published model of the death domain: Q29, E92, F93, F94 at the surface of helix 2 and helix 5 (37). Furthermore, residues which potentially interact with IRAK-1 in the N-termini and the end of the helix 4: R12, C13, E69, D73, T76 were also predicted using PPI-Pred (36). T66 in the death domain of IRAK-1 is critical for interaction with signaling molecules reported by Neumann and coworkers (38). In our model of the IRAK-M death domain, the following residues in the homologous positions were found: C35, R47, E59, D63, T69, Q78, D85, R96, R97 respectively. Taking into account their flexibility and accessibility (Fig.1A), the residues, Q78, F18, D19, P21, P22, A23, W74, L20, R70, S75, R97, Y105, A77, L101, in order of likelihood, were predicted as potential sites of interaction with IRAK-M ligands.

Combined with residue fluctuations as determined from MD simulation, hydrogen bond network analysis and homologue analysis, we proposed the mutagenesis of selected residues which we hypothesized to be involved in protein-protein interaction (Fig.1C).

The formation of IRAK-M-DD tetramer and the preparation of IRAK-4-DD tetramer, IRAK-2-DD tetramer and the complex of IRAK-2–IRAK-4–MyD88

Four IRAK-M-DD modules were superimposed to either IRAK-2 (3D packing quality: -1.231) or IRAK-4 (3D packing quality: -1.241) DD tetramer respectively in the crystallographic model (3MOP, (3)) by utilizing ICM-Pro Molsoft. The coordinates of four superimposed IRAK-M-DD were merged into one coordinate of IRAK-M-DD tetramer. The two obtained IRAK-M-DD tetramers were highly similar (Fig. 15) with an all atom RMSD value of 1.05 Å.

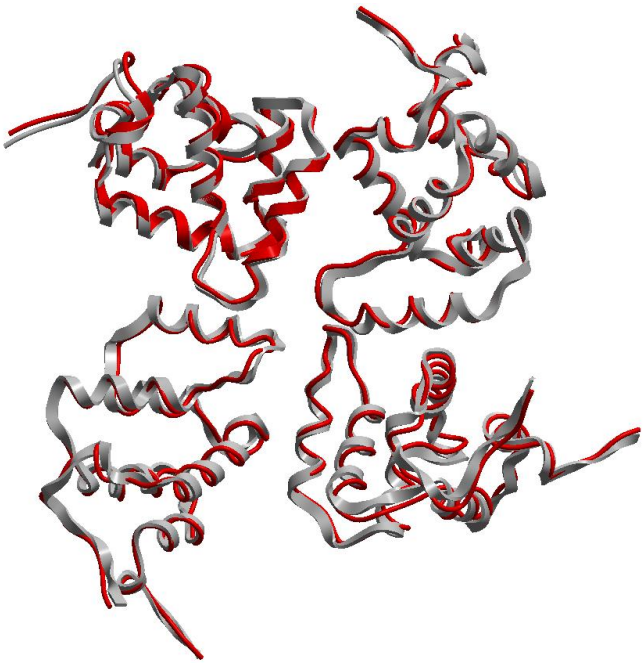


Figure 15. Superposition of the IRAK-M-DD tetramers based on IRAK2-DD tetramer in 3MOP (gray) and IRAK-4-DD tetramer in 3MOP (Black). The 3D structure of those tetramers were highly similar and the all atom backbone root mean standard deviation (RMSD) between them was 1.05 Å.

With the slightly higher 3D packing quality we used the IRAK-2-DD derived tetramer for further protein-protein docking studies. We first analyzed the atomic interactions (contact distance <=4 Å) between

the individual IRAK-M monomers in the IRAK-M-DD tetramer (Fig. 5). Next, IRAK-4 tetramer, IRAK-2 tetramers and IRAK-2–IRAK-4–MyD88 complex were retrieved from the crystallographic model 3MOP (3), with all hydrogen atoms added according to the protonation state at pH=7, which means that all acidic residues (Asp and Glu) are deprotonated and all basic residues (Lys and Arg) are protonated.

Protein-protein docking

Hex protein docking (39) was applied to dock the IRAK-M-DD tetramer to IRAK-4-DD tetramer, IRAK-2-DD tetramer and IRAK-2–IRAK-4–MyD88 complex respectively. We analyzed 100 docking events for IRAK-4-DD tetramer binding to IRAK-M-DD and found that 63% of the docked poses were to the W74 exposing side of the IRAK-M-DD tetramer and 37% to the R97 exposing side of the IRAK-4-DD tetramer. The IRAK-M-IRAK4-DD complex was then docked with another IRAK-M-DD tetramer and the docked results showed that the sandwich IRAK-M-IRAK-4-IRAK-M complex was formed, where the W74 exposing site of IRAK-M tetramer docked with the top surface (IRAK-2-tetramer binding surface in 3MOP) of IRAK-4-tetramer and the R97 exposing site of IRAK-M tetramer docked with the bottom surface (MyD88 binding surface in 3MOP) of IRAK-4 tetramer.

Docking of the IRAK-M-DD tetramer to the IRAK-2-DD tetramer indicated that in the vast majority (83%) this would lead to a W74 exposing side (IRAK-M-DD tetramer) interaction and only in 17% in

an interaction with the R97 exposing side. The IRAK-M-DD tetramer was docked to the top surface (consistent with the nomenclature in 3MOP (3) of the IRAK-2-DD tetramer from the IRAK-2–IRAK-4–MyD88 complex (Fig. 6B). The atomic interactions (contact distance <=4 Å) between IRAK-M-DD tetramer and the 3MOP structure were analyzed and three main binding areas were identified in green, pink and black (Fig. 6B).

Mutagenesis and expression of human IRAK-M mutants

Human IRAK-M and IRAK-1 mammalian expression vectors (pUNO) containing a blasticidin resistance element and the HA-TRAF6 vector were obtained from Invivogen (Toulouse, France). IRAK-M mutants were generated by site-directed mutagenesis with the QuickChange kit (Stratagene, La Jolla, CA) as recommended by the manufacturer with primers containing the desired mutations. Creation of an IRAK-M mutant which lacks the entire Death Domain was accomplished by introduction of a second SacII site (aaacta -> ccgcgg) in the ORF at codon position 103-105. SacII restriction of this mutated plasmid releases a fragment by cleavage at the introduced site and at the endogenous SacII site in codon 7-9 of the ORF. Ligation of the plasmid generates an ORF that encodes for amino acid 1-9 connected to 105-596 without introduction of additional amino acids. Constructs were subjected to DNA sequencing to confirm the mutations and check of the appropriate CDS. 293T cells, maintained in DMEM containing 10% FCS, were transfected using Lipofectamine 2000 (Invitrogen) as recommended by the manufacturer with the plasmids for protein expression and induction of NF-κB activation

and IL-8 release essentially as described (1). Lentiviral expression was obtained by introduction of WT IRAK-M in the pHEF vector and virus production in 293T cells as described (40, 41). The different IRAK-M Death Domain mutations were introduced in the constructed lentiviral human IRAK-M / IRES-eGFP expression system by restriction digestion of the pUNO-IRAK-M mutants with AgeI and BstB1 to obtain the mutated region which was ligated in the pHEF-IRAK-M vector which was digested with Xma1 (compatible overhang with Age1) and BstB1 to open up and remove the WT sequence. Non-replicating lentivirus production for transduction and expression of these mutants in cells was performed according to standard procedures (40). Cell lines were transduced with the generated lentiviral constructs by standard procedures (41) with addition of polybrene. After lentivral transduction with only eGFP as control, WT IRAK-M or the mutants, the populations of cells expressing the transgene were selected by cell sorting using the GFP signal as described (26).

The monocyte cell line THP-1 and transduced cultures were maintained in RPMI with 10% FCS and pen/strep matured to adherent macrophages and stimulated as described under conditions that will induce LPS tolerance after a previous exposure to LPS (42). The bronchial type lung epithelial cell line H292, which is devoid of IRAK-M transcripts, was maintained as well as the transduced cultures in IMDM with 10% FCS and pen/strep.

Western blotting

Westernblotting for IRAK-M was performed with 1 µg/ml monospecific rabbit anti-human IRAK-M antibodies directed against the C-terminal (Cell Signaling) or polyclonal mouse anti-full length human IRAK-M (Abnova) essentially as described (20). Westernblotting for phospho-ERK1/2 was performed using 1 µg/ml rabbit anti-pERK-1/2 (Cell Signaling). Western blotting for HA-TRAF6 was performed with 1 µg/ml anti-HA (Thermo Scientific).

NF-κB activation

Activation of NF-κB was determined 24 hours after transfection by cotransfection of a Firefly Luciferase NF-κB driven reporter construct and a Renilla Luciferase CMV driven construct on cell samples lysed with lysis buffer supplied with the DualGlo kit (Promega) used to determine the Firefly and Renilla luciferase activity in the same sample as recommended by the manufacturer.

IL-8 expression

For IL-8 expression by transfected 293T cells, the cells were washed 24 hours after transfection and fresh culture medium was placed on the cells and supernatant was harvested 24 hours later and stored at -20^{o}C for ELISA. Transduced H292 cells were plated in 96-wells cell culture plates (50.000 cells/well) and grown to confluency in 3 days and the medium was refreshed 24 hours before stimulation. Immediately before stimulation cells were washed again with medium again and cells were stimulated with 2 ng/ml IL-1 (Miltenyi Biotec) or 25 ng/ml Flagellin (Invivogen) for 6 hours and supernatant was collected and stored at -20^{o}C for ELISA. IL-8 was determined using

the duoset capture and biotinylated detecting antibody as recommended by the manufacturer (Invitrogen).

TNF and IL-6

The monocyte cell line THP-1 and transduced cultures were maintained in RPMI with 10% FCS and pen/strep matured to adherent macrophages and stimulated as described under conditions that will induce LPS tolerance after a previous exposure to LPS (39). Supernatant of matured THP-1 cells stimulated with 1 ng/ml LPS (ultrapure, Invivogen) or 500 ng/ml PAM3CSK4 (Invivogen) for 6 hour was collected and stored at -20^{o}C for determination of TNF and IL-6 by CBA (BD Biosciences) as described (43).

A20, SHIP and SOCS-3 expression

Negative regulators of inflammation were determined at the mRNA level by RT-qPCR as described (20).

Table 1. Predicted effect of the mutations in IRAK-M death domain.

Residue substitution	Change in solvent accessibility	Pseudo ΔΔ G (kcal mol-1)	Predicted effect on protein stability
F18A	-12.10%	2.36	No effect
D19N-L20A-P21A	38.7%	-0.5	No effect
P22A-A23S	14.6%	2.16	No effect
R70Q	-2.4%	-0.26	No effect
W74A	-12.1%	2.36	No effect
Q78A	4.3%	0.99	No effect

F18A-P21A	26.6%	1.86	No effect
D19N-A23S	53.3%	1.66	No effect
F18A/Q78A	-8.1%	3.35	No effect
R97Q	-8.3%	-0.07	No effect
Y105A	-1.6%	2.22	No effect
R97Q/Y105A	-9.9%	2.15	No effect

Predicted effect of IRAK-M death domain mutations on structural stability.
The mutations studied in the present include 11 individual residues: F18, D19, L20, P21, P22, A23, R70, W74, Q78, R97, and Y105. The 3D model of the wild type IRAK-M death domain was *in silico* mutated at these 11 residues by the YASARA/WhatIf twinpackage, followed by a 3 nanoseconds molecular dynamic simulation with the yasara2 force field in water to optimize the structure. The mutated model structures were next evaluated by the online server SDM (http://www-cryst.bioc.cam.ac.uk/~sdm/sdm.php) to predict the structural stability by the mutation [49]. This method applies a statistical potential energy function to calculate the pseudo ΔΔ free energy by using properties such as environment-specific amino acid substitution frequencies from the targeted protein homologous families. This value is comparable to the free energy difference between wild type and the mutant. From the server, five parameters: secondary structure of the mutated residue, solvent accessibility changes, hydrogen bond changes, pseudo ΔΔG and predicted effect on protein stability are generated for each single amino acid mutation. The solvent accessibility changes were the difference of the solvent accessibility (%) of the mutated residue and the wild type residue. F18A and W74A lost around 12% of the solvent accessible area and highly stabilized the structures while D19N and P21A gained 17.7% and 21.2% of the solvent accessible area.

The predicted effect of the multiple mutations on structural stability and the change in solvent accessibility of multiple mutations were calculated by the sum of the values from corresponding individual mutation values. The mutants with which experiments were performed in our work are shown in the table. These mutants were predicted to stabilize or did not affect the death domain structure. Meanwhile, the multiple mutants D19N-L20A-P21A, P22A-A23S, F18A-P21A, D19N-A23S increased the solvent accessible area.

Table 2. The structural quality of the template (2A9I from PDB_REDO (32)) used to build human IRAK_M model.

R Value	0.17	Backbone conformation[1]	-0.82
1st packing[1]	-0.81	Bond length RMS Z-score[2]	0.48
2nd packing[1]	-1.02	Bond angle RMS Z-score[2]	0.70
Ramachandran	0.82	Total number of bumps[3]	8
Chi-1/Chi-2[1]	-0.18	Unsatisfied H-bond donors/acceptors[3]	3

[1] Higher is better; [2] Should be lower than 1; [3] Fewer is better; Full WHAT_CHECK results of the template can be viewed in http://www.cmbi.ru.nl/pdb_redo/a9/2a9i/wf/index.html.

Table 3. Quality check of IRAK-M-DD model and the template by programs PROCHECK, WHATIF, VERIFY-3D, ERRAT and PROVE.

	IRAK-M-DD Model	Template (2a9i)
RMSD (Å)[1]	1.88	
3D Packing Quality[2]	-0.38	-0.07
Surface Area (Å^2)	7037.9	6987
Ramachandran abnormal	L39	-
Chi1-Chi2 abnormal	Y61 W76	L60
Planarity abnormal	W41	D73
Bump	G65-K66	-
1D-3D consensus[2]	87.38%	96.19%
Errat[2]	99.00%	100.00%
PROVE z-score[2]	0.4 (-5, 5)	0.153 (-5, 5)

[1] The backbone root mean square distance between IRAK-M-DD and the template. [2] The higher the better. Errat is to analyze the non-bonded interactions in protein 3D structures and generate confidence limits (0-1) to judge reliability of a protein's 3D structure [47]. PROVE is to check the atomic volume [48] and calculate the RMSD z-score between a given protein structure and the PDB dataset. The ideal PROVE z-score is expected as 0. A negative z-score means the atom volume smaller than average while a positive z-score means the atom volume greater than average.

Acknowledgements

This work was supported by grants from the Transnational University Limburg (to GN), and by a grant from the China Scholarship Council to Jiangfeng Du (grant no. 2008630114).

References

1. Wesche H*, et al.* (1999) IRAK-M is a novel member of the Pelle/interleukin-1 receptor-associated kinase (IRAK) family. *J Biol Chem* 274(27):19403-19410.
2. Muzio M, Ni J, Feng P, & Dixit VM (1997) IRAK (Pelle) family member IRAK-2 and MyD88 as proximal mediators of IL-1 signaling. *Science* 278(5343):1612-1615.
3. Lin SC, Lo YC, & Wu H (2010) Helical assembly in the MyD88-IRAK4-IRAK2 complex in TLR/IL-1R signalling. *Nature* 465(7300):885-890.
4. Nagpal K*, et al.* (2011) Natural loss-of-function mutation of myeloid differentiation protein 88 disrupts its ability to form Myddosomes. *J Biol Chem* 286(13):11875-11882.
5. Takeuchi O & Akira S (2010) Pattern recognition receptors and inflammation. *Cell* 140(6):805-820.
6. Darnay BG, Ni J, Moore PA, & Aggarwal BB (1999) Activation of NF-kappaB by RANK requires tumor necrosis factor receptor-associated factor (TRAF) 6 and NF-kappaB-inducing kinase. Identification of a novel TRAF6 interaction motif. *J Biol Chem* 274(12):7724-7731.
7. Ye H*, et al.* (2002) Distinct molecular mechanism for initiating TRAF6 signalling. *Nature* 418(6896):443-447.
8. Wan Y*, et al.* (2009) Interleukin-1 receptor-associated kinase 2 is critical for lipopolysaccharide-mediated post-transcriptional control. *J Biol Chem* 284(16):10367-10375.
9. Zhou H*, et al.* (2013) IRAK-M mediates Toll-like receptor/IL-1R-induced NFkappaB activation and cytokine production. *EMBO J* 32(4):583-596.
10. Flannery S & Bowie AG (2010) The interleukin-1 receptor-associated kinases: critical regulators of innate immune signalling. *Biochem Pharmacol* 80(12):1981-1991.
11. Kobayashi K*, et al.* (2002) IRAK-M is a negative regulator of Toll-like receptor signaling. *Cell* 110(2):191-202.
12. Su J, Xie Q, Wilson I, & Li L (2007) Differential regulation and role of interleukin-1 receptor associated kinase-M in innate immunity signaling. *Cell Signal* 19(7):1596-1601.

13. Su J, Zhang T, Tyson J, & Li L (2009) The interleukin-1 receptor-associated kinase M selectively inhibits the alternative, instead of the classical NFkappaB pathway. *J Innate Immun* 1(2):164-174.

14. Deng JC*, et al.* (2006) Sepsis-induced suppression of lung innate immunity is mediated by IRAK-M. *J Clin Invest* 116(9):2532-2542.

15. Hoogerwerf JJ*, et al.* (2012) Interleukin-1 receptor-associated kinase M-deficient mice demonstrate an improved host defense during Gram-negative pneumonia. *Mol Med* 18:1067-1075.

16. van der Windt GJ*, et al.* (2012) Interleukin 1 receptor-associated kinase m impairs host defense during pneumococcal pneumonia. *J Infect Dis* 205(12):1849-1857.

17. Xie Q, Gan L, Wang J, Wilson I, & Li L (2007) Loss of the innate immunity negative regulator IRAK-M leads to enhanced host immune defense against tumor growth. *Mol Immunol* 44(14):3453-3461.

18. Hubbard LL*, et al.* (2010) A role for IL-1 receptor-associated kinase-M in prostaglandin E2-induced immunosuppression post-bone marrow transplantation. *J Immunol* 184(11):6299-6308.

19. Balaci L*, et al.* (2007) IRAK-M is involved in the pathogenesis of early-onset persistent asthma. *Am J Hum Genet* 80(6):1103-1114.

20. van 't Veer C*, et al.* (2007) Induction of IRAK-M is associated with lipopolysaccharide tolerance in a human endotoxemia model. *J Immunol* 179(10):7110-7120.

21. Lasker MV, Gajjar MM, & Nair SK (2005) Cutting edge: molecular structure of the IL-1R-associated kinase-4 death domain and its implications for TLR signaling. *J Immunol* 175(7):4175-4179.

22. Ellinger-Ziegelbauer H, Brown K, Kelly K, & Siebenlist U (1997) Direct activation of the stress-activated protein kinase (SAPK) and extracellular signal-regulated protein kinase (ERK) pathways by an inducible mitogen-activated protein Kinase/ERK kinase kinase 3 (MEKK) derivative. *J Biol Chem* 272(5):2668-2674.

23. Ioannidis I, Ye F, McNally B, Willette M, & Flano E (2013) Toll-like receptor expression and induction of type I and type III interferons in primary airway epithelial cells. *J Virol* 87(6):3261-3270.

24. Paplinska-Goryca M, Nejman-Gryz P, Chazan R, & Grubek-Jaworska H (2013) The expression of the eotaxins IL-6 and CXCL8 in human epithelial cells from various levels of the respiratory tract. *Cell Mol Biol Lett* 18(4):612-630.

25. Yu Y*, et al.* (2003) TLR5-mediated activation of p38 MAPK regulates epithelial IL-8 expression via posttranscriptional mechanism. *Am J Physiol Gastrointest Liver Physiol* 285(2):G282-290.

26. Kobayashi H*, et al.* (2011) Neutrophils activate alveolar macrophages by producing caspase-6-mediated cleavage of IL-1 receptor-associated kinase-M. *J Immunol* 186(1):403-410.

27. Nguyen T, De Nardo D, Masendycz P, Hamilton JA, & Scholz GM (2009) Regulation of IRAK-1 activation by its C-terminal domain. *Cell Signal* 21(5):719-726.

28. Burns K*, et al.* (2003) Inhibition of interleukin 1 receptor/Toll-like receptor signaling through the alternatively spliced, short form of MyD88 is due to its failure to recruit IRAK-4. *J Exp Med* 197(2):263-268.

29. Li X, Commane M, Jiang Z, & Stark GR (2001) IL-1-induced NFkappa B and c-Jun N-terminal kinase (JNK) activation diverge at IL-1 receptor-associated kinase (IRAK). *Proc Natl Acad Sci U S A* 98(8):4461-4465.

30. Westbrook J, Feng Z, Chen L, Yang H, & Berman HM (2003) The Protein Data Bank and structural genomics. *Nucleic Acids Res* 31(1):489-491.

31. Abagyan R, Frishman D, & Argos P (1994) Recognition of distantly related proteins through energy calculations. *Proteins* 19(2):132-140.

32. Joosten K*, et al.* (2008) A knowledge-driven approach for crystallographic protein model completion. *Acta Crystallogr D Biol Crystallogr* 64(Pt 4):416-424.

33. Morozov AV & Kortemme T (2005) Potential functions for hydrogen bonds in protein structure prediction and design. *Adv Protein Chem* 72:1-38.

34. Fernandez-Recio J, Totrov M, Skorodumov C, & Abagyan R (2005) Optimal docking area: a new method for predicting protein-protein interaction sites. *Proteins* 58(1):134-143.

35. Chen H & Zhou HX (2005) Prediction of interface residues in protein-protein complexes by a consensus neural network method: test against NMR data. *Proteins* 61(1):21-35.

36. Bradford JR & Westhead DR (2005) Improved prediction of protein-protein binding sites using a support vector machines approach. *Bioinformatics* 21(8):1487-1494.

37. Mendoza-Barbera E*, et al.* (2009) Contribution of globular death domains and unstructured linkers to MyD88.IRAK-4 heterodimer formation: an explanation for the antagonistic activity of MyD88s. *Biochem Biophys Res Commun* 380(1):183-187.

38. Neumann D*, et al.* (2008) Threonine 66 in the death domain of IRAK-1 is critical for interaction with signaling molecules but is not a target site for autophosphorylation. *J Leukoc Biol* 84(3):807-813.

39. Ghoorah AW, Devignes MD, Smail-Tabbone M, & Ritchie DW (2013) Protein docking using case-based reasoning. *Proteins* .

40. Seppen J, Rijnberg M, Cooreman MP, & Oude Elferink RP (2002) Lentiviral vectors for efficient transduction of isolated primary quiescent hepatocytes. *J Hepatol* 36(4):459-465.

41. van Lent AU*, et al.* (2009) IL-7 enhances thymic human T cell development in "human immune system" Rag2-/-IL-2Rgammac-/- mice without affecting peripheral T cell homeostasis. *J Immunol* 183(12):7645-7655.

42. Martin M, Katz J, Vogel SN, & Michalek SM (2001) Differential induction of endotoxin tolerance by lipopolysaccharides derived from Porphyromonas gingivalis and Escherichia coli. *J Immunol* 167(9):5278-5285.

43. de Vos AF*, et al.* (2009) In vivo lipopolysaccharide exposure of human blood leukocytes induces cross-tolerance to multiple TLR ligands. *J Immunol* 183(1):533-542.

Chapter 5

Homology modeling and binding site prediction of human IRAK-M

Jiangfeng Du, Cornelis van 't Veer, Kanin Wichapong, Gerry A.F. Nicolaes

Submitted to Chinese Science Bulletin

Abstract

Human Interleukin-1 receptor-associated kinase M (IRAK-M) is a member of the IRAK protein family and is mainly present in the marrow tissue. IRAK-M is an important regulator of inflammation in the innate immune system. The function of IRAK-M is related to many diseases such as asthma. IRAK-M consists of two important domains: a death domain (DD) and a kinase domain (KD). Homology models of the death domain and kinase domain were built based on the crystallographic structures (PDBID: 2A9I, 2NRU) of its homolog IRAK-4, with sequence identities being: 28.7% and 30% respectively. In the death domain, residues F18, D19, L20, P21, P22, A23, R70, W74, S75, Q78, R97, Y105, A107 were predicted to be at the protein-protein interaction interface, based on analysis of the mouse IRAK-4 and protein-protein interaction prediction (PPI) programs ODA (Molsoft), Cons-PPISP and PPI-Pred, combined with molecular dynamic properties such as residue flexibility and hydrogen bond networks. The identified residues form two distinct binding surfaces. From our model of the kinase domain of IRAK-M, we conclude that IRAK-M likely contains an inactive serine kinase domain that possesses several features that are common to kinase domains but essentially the IRAK-M kinase domain lacks a primary phosphorylation site. In the activation region (S333-Y340), serine residues have been substituted by threonine residues, which may influence the geometry of the activation loop. Moreover, the ATP binding pocket is narrow and does not have an efficient solvent shield as compared to its template IRAK-4 kinase domain. IRAK-M

residues L210, E214, L217, M314, H316, C326, T327 are predicted to be in direct interaction with antagonists or substrates such as ATP.

The atomistic models represent possible 3D structures of the IRAK-M death domain and kinase domain. These structures can be used to decipher the structure function relationships of IRAK-M and may facilitate rational drug discovery to regulate the expression of IRAK-M activity.

Keywords IRAK-M, Homology Model, Kinase, MD simulation, protein-protein interaction (PPI)

Introduction

Interleukin-1 receptor-associated kinases (IRAK) are intracellular signalling proteins involved in the innate immune system. The first described member of this family, IRAK-1, was described as a signal transducer for interleukin-1 (IL-1) (1). IRAK-mediated signaling is initiated through activation of a receptor from the interleukin-1 receptor / Toll-like receptor superfamily, which in turn triggers the recruitment of myeloid differentiation primary response protein (MyD88) and IRAK complexes. When complexed, IRAK-4 can phosphorylate IRAK-1, which then results in recruitment of TNF-association factor 6 (TRAF6), upon which through a series of signaling events NFκB is induced. (2). The IRAK family of protein kinases has four different members: IRAK-1, IRAK-2, IRAK-3 and IRAK-4. Both IRAK-1 and IRAK-2 can bind to the Toll-like receptor adaptor protein MyD88 and trigger intracellular signaling cascades that lead to transcriptional up-regulation and mRNA stabilization. IRAK-4 forms a complex with IRAK1, in a reaction that is most efficient when both IRAK-4 and IRAK1 are bound to MyD88, upon which IRAK-4 phosphorylates IRAK1. IRAK-3, also called IRAK-M, inhibits the dissociation of IRAK1 and IRAK4 from the Toll-like receptor complex through an as yet unclear mechanism, but which may involve inhibition of phosphorylation of IRAK1 and IRAK4 or stabilization of the complex between the receptor, IRAK-1 and IRAK-4. IRAK-1 and IRAK-4 contain active kinase subunits whereas IRAK-2 and IRAK-M do not possess kinase activity. All IRAK members mediate activation of nuclear factor-kB (NFκB) and mitogen-activated protein kinase (MAPK) via the TLR/IL-1 pathway (3). The signal

transduction pathways that are initiated by toll/IL-1 receptor family members ultimately lead to the activation of transcription factors such as activator protein 1 (AP-1) or NFκB, which on their turn contribute to the establishment of an immune response.

The different IRAK family members have varying lengths: IRAK-1 is 693 amino acids in length, IRAK2 625, IRAK-3 596, and IRAK-4 460. All IRAK family members have homologous overall multi-domain protein architecture, consisting of a conserved N-terminal death domain (DD) and a central kinase domain (KD) (1, 2, 4-6). The DD is a protein interaction motif, implicated in binding of IRAK proteins to the adaptor protein MyD88 (4, 7). Kinase domains contain an active site responsible for phosphorylation of specific substrates and a binding site for ATP. Amongst the IRAK family members, all forms have a functional ATP binding pocket with a conserved lysine (K239 for IRAK1; K237 for IRAK2; K205 for IRAK3; K213 for IRAK4) as binding site in the kinase domain; only IRAK-1 and IRAK-4 contain a functional catalytic site that holds a critical aspartate residue (D340 for IRAK-1, D311 for IRAK-4) (1, 8, 9). In IRAK-2 and IRAK-M this critical residue has changed into an asparagine or a serine, respectively, which renders their kinase domain inactive (2). However, further understanding of the non-functional IRAK-M kinase domain is needed. In between the DD and the KD is a region of unknown function that has a low sequence similarity when compared between different IRAK family members (UD as an abbreviation for this unknown domain). The UD of IRAK-1 has been demonstrated to be essential for IL-1-mediated induction of NFκB (10). The UD is rich in prolines, serines, and threonines and contains two so-called PEST

sequences (Proline-Glutamate-Serine-Threonine), which could be involved in the degradation of IRAK-1 upon IL-1 stimulation (11). IRAK-2 lacks a PEST motif in the UD, which may increase its stability upon IL-1 treatment, although this has not been investigated yet (9). IRAK-1, IRAK-2 and IRAK-M but not IRAK-4 further contain a C-terminal stretch, which does not show any similarity to known protein motifs. (12, 13).

Wesche and co-workers showed that IRAK-M can heterodimerize with IRAK-1 (2). Moreover, IRAK-M was found to perform a role as negative regulator in TLR/IL-1R signaling pathway, as was concluded from experiments with IRAK-M deficient mice that exhibit increased TLR/IL-1R signaling(14). Exactly, how IRAK-M inhibits TLR/IL-1R signaling is still speculative, but it was shown that IRAK-M enhances the binding of MyD88 to IRAK-1 and IRAK-4 and prevents IRAK-1 phosphorylation. In this way, IRAK-M traps both IRAK molecules in the receptor complex, preventing the complex from dissociation.

A detailed three-dimensional structure of atomistic detail is presently not available for IRAK-M. Neither X-ray nor NMR structures have been published so far. In an effort to facilitate the rational design of structure-function experiments we have employed comparative modeling techniques to obtain a three-dimensional model for IRAK-M. As template structures we selected structures of IRAK family members that have been determined via X-ray crystallography or NMR spectroscopy. With these models we will be able to understand the intricate structure-function relationships of this molecule and rationalized targeted mutagenesis of IRAK-M will become possible.

Furthermore, if these structures contain druggable pockets, structure-based drug design should become a reality to interfere with the activities of IRAK-M.

Materials and Methods

Template Selection and Alignment

The primary sequence of IRAK-M consists of 596 amino acids. The exact definition of domain boundaries is however not unequivocal as different sources (Uniprot, NCBI, HPRD) define different domain boundaries. Thus, in order to include the complete domain region, we applied multiple sequence analysis of IRAK-M from different species (Wild Boar, Bovine, Dog, Human, Brown rat, Mouse, Zebrafish) (Figure 1) to verify the domain boundaries. The sequence used for modeling of the death domain was selected from C5 to G119 in this work. For the kinase domain, the sequence was chosen from E141 to L460. The potential templates for the IRAK-M death domain and kinase domain were selected through PSI-BLAST search of the protein database (PDB) (15) and the best template was selected as the final template, as judged by the sequence alignment score and structure quality (X-ray resolution and R-factor). The alignment between template and the IRAK-M sequence was performed by ICM-Pro (MolSoft) (16) with default scoring parameters and refined based on secondary structure, amino acid features and the 3D structure of the template (Figure 2).

Initial death domain model construction

The region from C5 to G119 in IRAK-M was selected as a query sequence to search for a potential template in the protein data Bank (PDB) (17). PSI-BLAST (PSSM: 0.005) results showed that there are 12 coordinate files available, with sequence identities between 24% and 30% with the IRAK-M death domain sequence, which may qualify as a template structure. We chose 2A9I as a template, based on its sequence identity (28.7%), which was also evaluated by the multiple sequence alignments; moreover, 2A9I describes the crystal structure of the death domain of IRAK-4 at a high resolution at 1.7 Å, for which an optimized coordination file was available from the PDB_REDO server (18) with 0.619 as the Z-score of the 3D packing quality.

Initial kinase domain construction

A number of X-ray structures of protein kinase domains can be used for comparative modeling of the IRAK-M kinase domain, these are PDBID: 2OIB, 2OIC, 2OID, 2O8Y, 2NRU, 2NRY. These coordinate files describe the human IRAK-4 kinase domain deposited in the PDB as a tetrameric structure in different conditions such as in apo form or with various substrates/inhibitors. Since all these structures represent experimentally determined X-ray structures, they share the same sequence length of 304 amino acids, identical BLAST scores and equal sequence identity with the kinase domain sequence of hIRAK-M. Taking into account the quality of the structures, as judged by their resolution, R-Value, B-factor and bond properties, we selected 2NRU.pdb as the best template. The sequence alignment between the 2NRU sequence and the kinase domain of IRAK-M

(residues E141 to L460) shows a sequence identity of 30% and reveals four gaps which exist in loop regions of the 2NRU structure (Figure 2b).

Figure 1. Multiple Sequence alignment of IRAK-M orthologs from "*Wild Boar, Bovine, Dog, Human, Brown rat, Mouse, Zebrafish*" by ICM-Pro package. The sequence of human IRAK-M was in green shade. The regions of the death domain and kinase domain were indicated by red and blue boxes respectively.

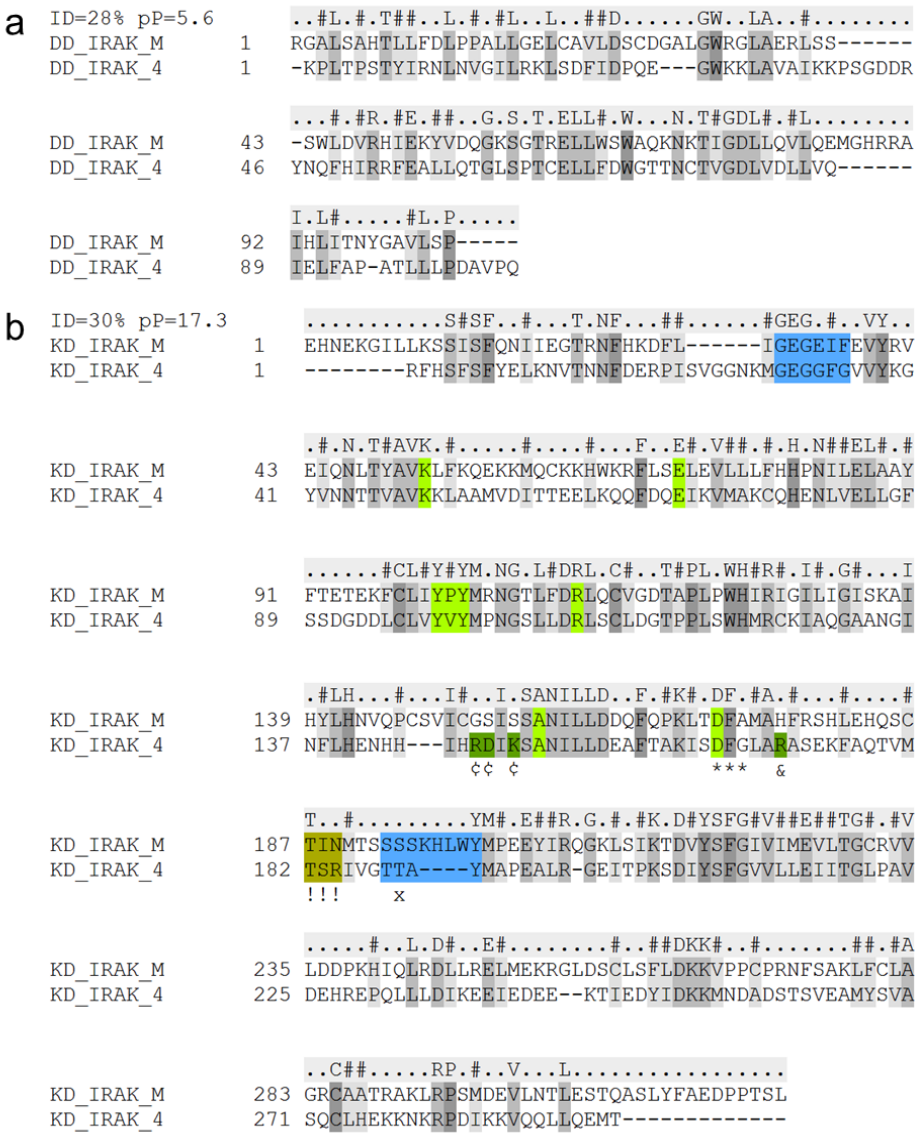


Figure 2. Sequence alignment between the human IRAK-M death domain and kinase domain and their templates. Fig. 2a, the death domains of human IRAK-M (residue R8-P111) and mouse IRAK-4, the amino sequence identity was 28.7%. Fig. 2b, the kinase domains of human IRAK-M and human IRAK-4, which have a 30% amino acid sequence identity. Residues K192, E212, Y241, P242, Y243, R252, A297 and D311 surrounding the ATP pocket are shown in light green. The activation loop and the P-loop were shown in light blue. The DGF motif was indicated by '*' in IRAK-4 but DGA was in the IRAK-M kinase domain. The primary

phosphorylation site was indicated by '!'. The critical threonine T352 is indicated by 'X' in IRAK-4. The important residues for substrate binding are shown in dark green shade and indicated by '¢. Multiple serine residues are indicated by 'X' in IRAK-M.

Unstructured Domain (UD) and C-terminus Region

UD ranges from G120 to P140 and the C-terminal region is defined from R464 to E596. No model structure could be obtained for these two regions. This is due to the fact that these regions show either no or insufficient homology to other proteins of known 3D structure. Moreover, the alignment lengths were such that only incomplete models (the coverages were less than 60% of the whole regions) could be built for these IRAK-M segments. Collectively, too little information is available at present to allow a reliable 3D structure prediction for these parts of IRAK-M.

Backbone Generation, Loop Modeling and side chain Sampling

In the aligned regions, the ICM-Pro molecular modeling package (16) was utilized to build the backbone conformation of IRAK-M. In the loop regions, conformations were constructed by LOOPY (19), an energy-based *ab initio* prediction program, which is based on the algorithm of Random Tweak and Direct Tweak to filter steric clashes and evaluate structural qualities. Side chain conformations were firstly modelled by use of a rotamer library and for the non-identical residues, the side chain torsion angles were sampled by a biased probability Monte Carlo global optimization of the energy on the basis of surface properties, entropy, and electrostatics. The side chain conformations were further minimized by a steepest descent and simulated annealing minimization using the ICM Pro package

(Version 3.4.6).

The Death Domain

After refinement of the sequence alignment as described above, we used ICM Pro to build a model for the death domain of IRAK-M. The next step was loop modeling. The homologous positions of the loops at residues D30-W38 and at residues L48-W52 were surface exposed and we expected these to have flexible structures. Thus, we decided to use LOOPY to rebuild these loops in our model. Side chains were minimized by steepest descent method and simulated annealing minimization with a fixed backbone conformation and next optimized without any restraint using the ICM pro package (Version 3.4.6).

The Kinase Domain

The kinase domain of IRAK-M could only be modeled between residues L148 and Q447 because of limitations in the template structure. This sequence includes five loop regions, which are the fragments from F169-G172, V284-I290, S334-M341, I346-G349, and K393-S398 respectively. In analogy to the method applied for the death domain modeling, we utilized ICM Pro and LOOPY to build the kinase domain backbone structure and then side chains were minimized by steepest descent method and simulated annealing minimization with a fixed backbone conformation, next the resulting kinase structure was optimized by use of the ICM Pro package (Version 3.4.6).

Molecular dynamics (MD) simulation and Quality Check

Short MD simulations were performed to optimize the homology models by use of the Yasara-Whatif twin package (20, 21). The MD simulation systems contained explicit water (0.997 kg/L, pH 7.0), under the YAMBER3 force field, where the default Yamber3 cutoff for long range Coulomb interaction of 7.86Å was chosen and we employed periodic boundaries for the simulation boxes. The frequency of the simulation trajectory is 25 picoseconds (ps), and every snapshot from the trajectory was submitted for online structure quality check at the NIH institute (http://nihserver.mbi.ucla.edu/-SAVES_3/), using PROCHECK (22), WHATIF (21), VERIFY-3D (23, 24), ERRAT (24), PROVE (25). The best snapshot in terms of the total energy was selected as the starting point for a long MD simulation (100 nanoseconds (ns) for the death domain and 25 ns for the kinase domain), where AMBER03 force field, periodic boundaries and PME for electrostatics were used with the explicit water (0.997 kg/L, pH 7.0, NaCl 0.9%). The resulting energies, RMSD, residue flexibility, hydrogen bonds and Ramachandran plots from the MD simulation were calculated by YASARA scripts and the VMD (26) program.

Prediction of protein-protein contacts and mutations in IRAK-M

Being a death domain, the DD of human IRAK-M is able to interact with other death domains, such as those present in MyD88 and other IRAK protein family members. Identification and description of potential protein-protein interaction areas therefore could be helpful

to study and explain the functions of IRAK-M. A consensus approach was used to predict protein-protein interactions for the DD model structure generated as described above. We employed several methods to do so: Optimal Docking Area (ODA) (27), Cons-PPISP (28) and PPI-Pred (29). For each of these methods, a prediction score was obtained for all residues in either the death domain or the kinase domain of IRAK-M and a consensus was generated from the different methods applied, taking into account the accessibility of the residue.

Results and Discussion

Homology modeling of the IRAK-M death domain

We used PROCHECK, WHATIF, VERIFY-3D, ERRAT and PROVE to check the structural quality of the models that were generated for the death domain. Excluding the amino acids glycine and proline, 90% of residues (81 residues) were in the most favored zones after refinement and the remaining 8.9% (8 residues) located in allowed regions as analyzed by inspection of corresponding Ramachandran plot (Figure 3). The 3D packing quality was evaluated by WHATIF as -0.38 (acceptable range: (-2.5, 0)) (30), which indicates a good internal configuration of the model. The death domain presented a surface area of 7038 $Å^2$. The inner atomic distances were in agreement with the reference X-ray structures (PDB database) except for a bump in a loop region, between G65 and K66. The non-bonding interactions in the structure were qualified as “highly packed” by the ERRAT program (24). The high quality of the model was

supported by the residue environment assessments where 99% of all residues had statistically preferable non-bonded interactions. The atom volume quality was evaluated by PROVE (31) and showed the atom volume was slightly better than the averaged high resolution x-ray structures in PDB database.

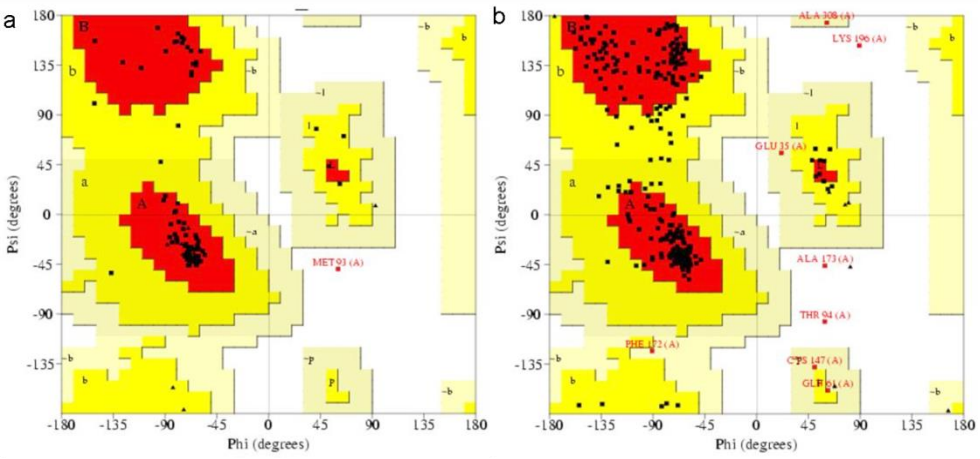


Figure 3. Ramachandram plots of the human IRAK-M death domain (Fig. 3a) and kinase domain (Fig. 3b). The black dots indicate the individual amino acids. The red areas represent favorable regions, while yellow areas represent the acceptable regions and the white region the theoretically forbidden region. M93, T234, H323 and K336 are being studied for any special structure function relationship.

The death domain of human IRAK-M (Figure 4a) folded into a protein-protein binding motif as is present in the structural template, mouse IRAK-4, with a 6 helical bundle forming a hydrophobic core that was decorated with a charged outer layer. K60 located at the C-terminus of helix3 while E71, in the N-terminus of helix4, made the helix dipole force favorable. The intrinsic dipole potential of helix1 influenced the side chain conformations of D36 and R42. In the IRAK-M DD model, a unique anti-parallel beta sheet was formed by

one strand from the N-terminus and another strand from the loop that precedes helix 5. An anti-parallel sheet located in between helix2 and helix3 in the template was replaced by a sharp turn (S50-S51) in our model. Two extra hydrogen bonds had been added into the model by flipping the histidine side chain (H57, H95).

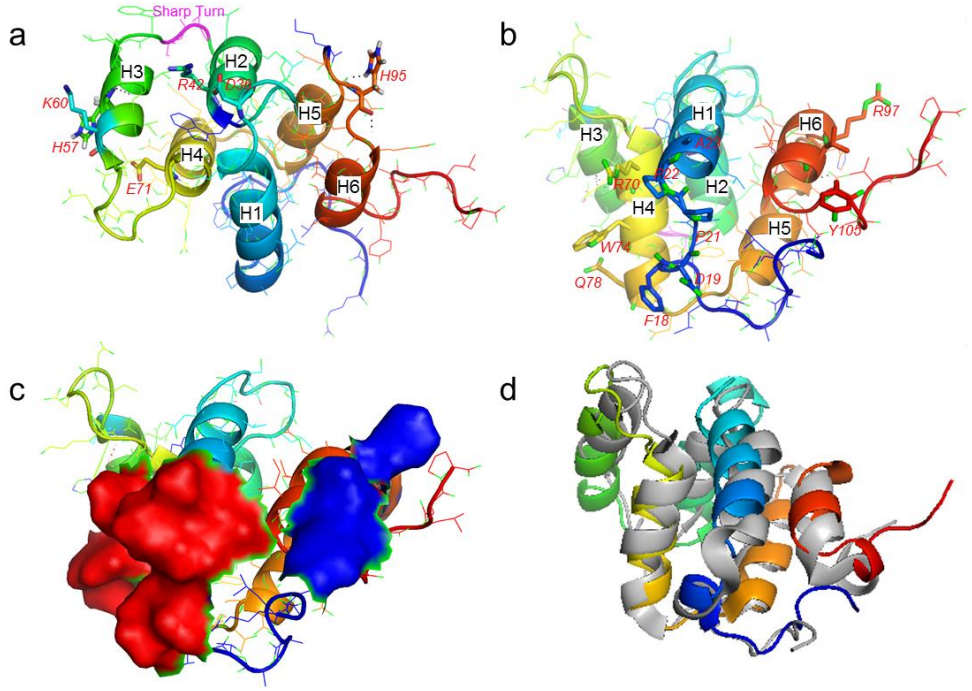


Figure 4. Homology model of human IRAK-M death domain. Fig. 4a, the death domain model was generated as described in the Methods section. The model was color-coded from N-terminus to C-terminus by Pymol (32). It contains 6 helices (H1-H6) and forms a typical beta barrel. A unique sharp turn is shown in magenta. H57 and H95 were flipped for optimizations. K60 and E71 enhance the helix dipole force (H3, H4). D36 and R42 are shown in the model. Fig. 4b, the residues which were predicted for protein-protein interaction are labeled in red and their side chains are solvent exposed. Fig. 4c, two interaction patches are shown in surface in the model (Patch 1: red and Patch 2: blue). Fig. 4d, Superimposition of the IRAK-M death domain model (rainbow) and its template (gray). The RMSD value of the aligned atoms is 1.088 Å.

A 100 ns MD simulation was performed to study the total energy, the

structural stability, hydrogen bond network and amino acid fluctuation (RSMF), of the IRAK-M death domain (Figure 5). During the MD simulation, the total energy was stable at -3.66 * 10^5 kJ/mol. Hydrogen bonds, known to be important for protein stability and function, were studied during the simulation (33). In the model, the number of the hydrogen bonds was stably kept as 19 during the simulation. The fluctuation of each residue during the MD simulation was calculated and we observed that the residues, which were predicted in PPI binding patches 1 and 2 (see also below), were stably presented to the solvent as the RMSF values in these patches were not very variable. In contrast, several loop regions, and in particular the loop (V62-K66) between helix3 and helix4 was flexible with N64 being most flexible. Residue M93 was analysed as being unfavorable in the Ramachandran check and it was observed that this residue appeared as one of the most flexible residues during the simulation.

The IRAK-M death domain is reportedly involved in signaling processes that ultimately lead to programmed cell death by apoptosis (34, 35) and in down-regulation of the NFκB pathway by prevention of the dissociation of IRAK1/IRAK4/MyD88 complex (14). The death domain represents a protein-protein interaction motif which allows self-association and association between different death domain motifs (Figure 6).

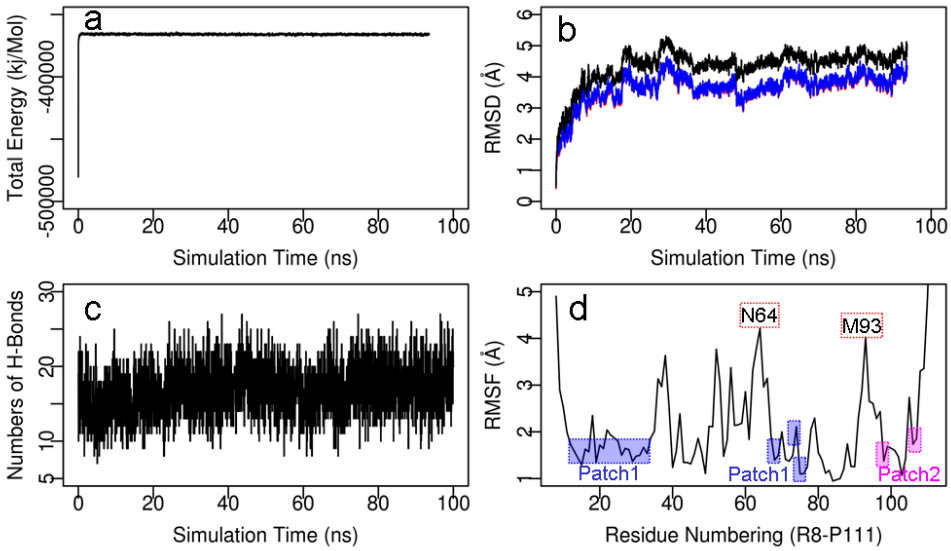


Figure 5. Properties of the human IRAK-M death domain model. Fig. 5a, total energy was monitored during the simulation. Fig. 5b, root mean square deviation (RMSD) of the alpha carbon (Ca) (Blue), backbone (orange) and heavy atom (black) were calculated during the simulation. As the RMSD values of the backbone atoms are nearly identical to the RMSD values of the Ca atoms, therefore the orange line cannot be observed clearly. Fig. 5c, the total number of hydrogen bonds in the domain along the MD simulation. Fig. 5d, the backbone RMS fluctuation of each residue in the domain was calculated, the x-axis represents the residue number from R8 to P111. The residues F18, D19, L20, P21, P22, A23, R70, W74, S75, A77, Q78, R97, L101 and Y105 which form the predicted interaction patches are shown in colored rectangles (Patch 1 in blue and patch 2 in magenta). Two most flexible residues (N64 and M93) are indicated by residue name and red box.

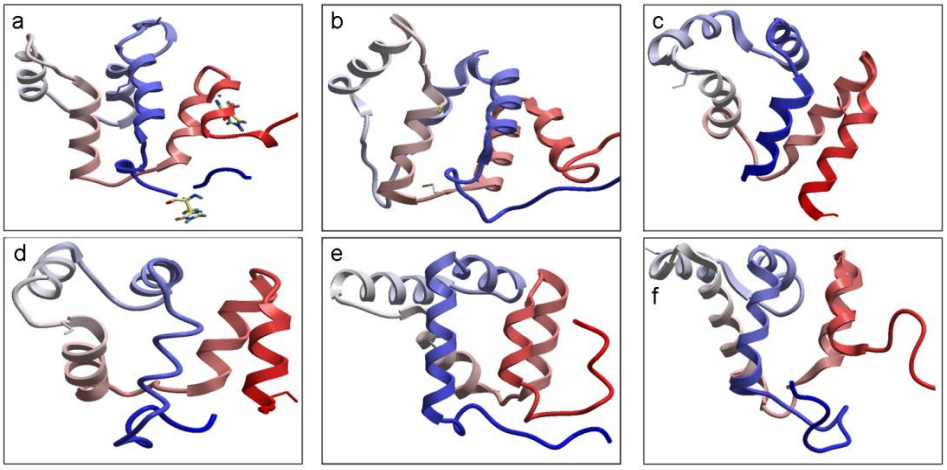


Figure 6. Comparison of the death domain motifs in different proteins. The motif is generally comprised of 6 helices and brings the N- and C-terminus in close proximity. Fig. 6a, human IRAK-M. Fig. 6b, mouse IRAK-4 (PDBID: 2A9I). Fig. 6c, human TN receptor (1ICH). Fig. 6d, rat neurotrophin receptor (1NGR). Fig. 6e, drosophila pelle (1YGO). Fig. 6f, human nuclear matrix P84 (1WXP).

In order to rationally design methods to interfere with death domain binding, it is helpful to predict which residues or regions are potentially involved in protein:protein interaction. Several important residues for the interaction between the IRAK-4 death domain and the MyD88 death domain have been confirmed on basis of an earlier published model of the death domain: Q33, E96, F97, F98 (36), which are at the outside of helix 2 and helix 5. Furthermore, residues which potentially interact with IRAK-1 at the N-terminus and at the C-terminus of the helix 4: R16, C17, E73, D77, T80 were also predicted using PPI-Pred (29).

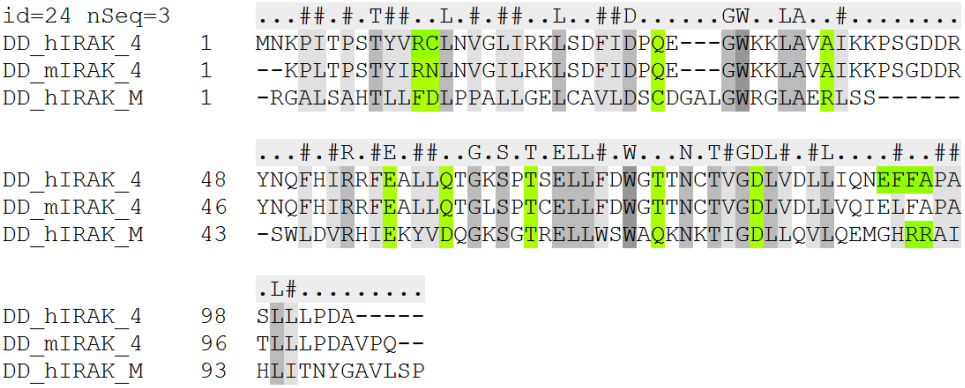


Figure 7. Multiple sequence alignment of the human IRAK-M death domain (DD), the human IRAK-4 DD and the mouse IRAK-4 DD. The residues, of which the homologous positions were confirmed as the protein-protein interaction regions in the IRAK-4 (29, 36), are shaded in light green.

T70 in the death domain of IRAK-1 is critical for interaction with signaling molecules, as was reported by Neumann and coworkers (37). In our model of the IRAK-M death domain, the following residues in positions, homologous to those mentioned for IRAK-4 and IRAK-1 were found: F18, D19, C35, R47, E59, D63, T69, Q78, D85, R96, R97 respectively (Figure 7). We employed a consensus protein-protein interaction prediction approach by application of several different methods to predict potential interaction residues: Optimal Docking Area (ODA) (27), Cons-PPISP (28) and PPI-Pred (29). The residues which were predicted by all the three programs were considered as potential sites for protein-protein interaction. These residues were further filtered by the residue flexibility derived from MD simulations (Figure 5d), the residues F18, D19, L20, P21, P22, A23, R70, W74, S75, A77, Q78, R97, Y105 were predicted as potential sites of interaction with IRAK-M ligands or modulators such as MyD88, IRAKs members. The selected residues are surface

exposed and do not form intra-molecular contacts (Figure 4b) by 3D structural analysis of our model structure.

The death domain involves the biological functions by directly binding to other death domains. Mutagenesis experiment on the predicted residues mentioned above may interrupt the death domain's functions. To guide a rational mutagenesis experiment, we applied a mutation prediction (38) based on a position specific rotamer and the prediction suggested that the mutations F18A, D19N, L20Q, P21A, P22N, A23K, R70Q, W74E, S75Y, Q78A, R97N, L101A Y105D do not impact the structural integrity and those mutated residues were predicted as the optimal solution, thus these mutations may be considered to study the function of the IRAK-M. Hydrogen bonds formed between N104 and A107 likely create an interaction between the C-terminal loop and helix 6. Breaking of the hydrogen bond probably results in a relief of conformational constraints for the C-terminal loop, which is likely to occur when a mutation at N104 is introduced. Residues identified here form two binding patches: one is formed by the N-terminus of helix1, the C-terminus of helix4 and the loop between helix4 and helix5, whereas R97N, I102S and Y105Q are likely important residues that form a second patch (patch 2) in helix6 (Figure 4c).

Homology modeling of the IRAK-M Kinase Domain

Similarly as was done for the DD model, the quality was checked for the KD. The best KD model (Figure 8a) was obtained after a 500 ps MD simulation and the resulting structure was next evaluated. 83.3%

of residues (240 amino acids) were located in the most favored zones and 13.9% (40 amino acids) in allowed regions, 1.7% (5 amino acids) in generally allowed regions and 1% (T234, H323 and K336) in disallowed regions (Figure 3b). Similar results were obtained for the template 2NRU: 88.0% of residues in the most favored zones, 9.8% in allowed regions, 1.9% in generally allowed regions, and 0.4% in disallowed regions.

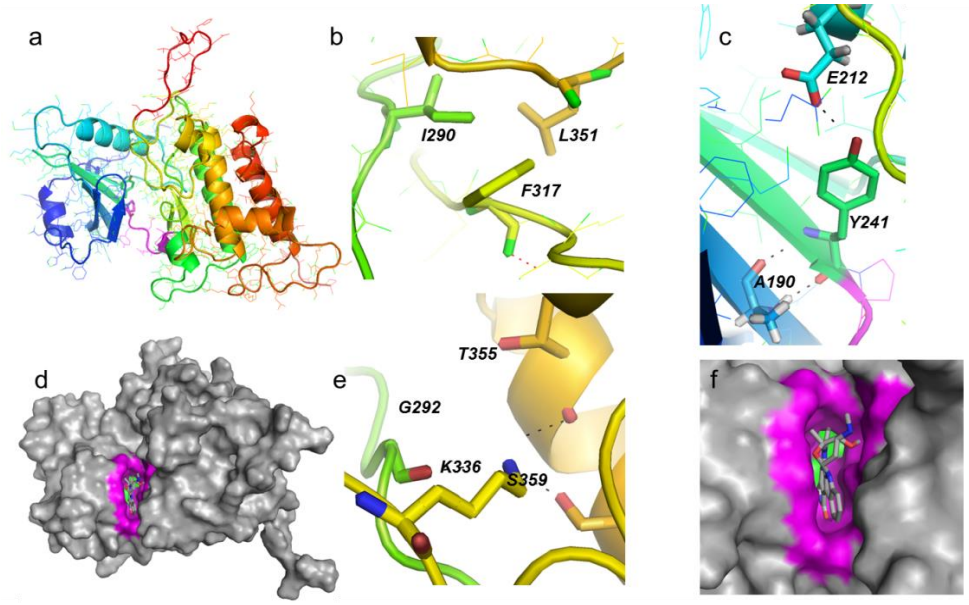


Figure 8. Homology model of the human IRAK-M kinase domain. Fig. 8a, the kinase model was generated by use of the ICM-pro package, combined with LOOPY. The model is color-coded from N-terminus to C-terminus by Pymol software. It consists of two subunits which are linked by a hinge region at P242-L249 (shown in magenta). The loop which is unique for IRAK-M is shown in red. Fig. 8b, the side chain was restricted by two ternary neighboring residues I290 and L351. Fig. 8c, the so-called gatekeeper residue Y241 forms three hydrogen bonds with A190 and E212. Fig. 8d, the proposed ATP binding pocket is shown in the presence of a pseudo substrate, an ATP molecule docked to the kinase model. Fig. 8e, K336 rendered three hydrogen bonds with G292, T355 and S359. Fig. 8f, an enlarged pocket from Fig. 8d, where the gatekeeper Y241 is shown in green and the surrounding residues are shown in magenta.

A number of common deviations from the ideal structures were found by WHATIF (21) analysis, both for the model and template structure. However, the side chain planarity of H282 and H326 were suggested to be optimized. F317 had an abnormal Chi1-Chi2 rotamer, which is caused by a repulsive force from I290 and L351 (Figure 8b), whereas the gatekeeper residue Y241 had an unfavorable side chain planarity in the model since the gatekeeper residue was pushed by a pseudo substrate (Figure 8c), an ATP molecule docked to the kinase model. We studied the relationship between the amino acid sequence and the 3D structure of the kinase model by application of the verify3D program (39) and we found that the predicted structure for 73.5% of all residues is in agreement with their 3D structures as obtained from the 3D model. The non-bonded interactions between different atom types in the kinase domain were 99.2% in agreement with those of the template structure, 2NRU. The main quality improvement achieved by performing the MD, as indicated by the results of PROVE, is in the volumes of atoms. There were no steric clashes in the model, which indicates a high 3D packing quality of the model. A 25 ns MD simulation was performed to analyze the IRAK-M kinase domain model with respect to its structural stability, amino acid fluctuation and potential energy changes (Figure 9).

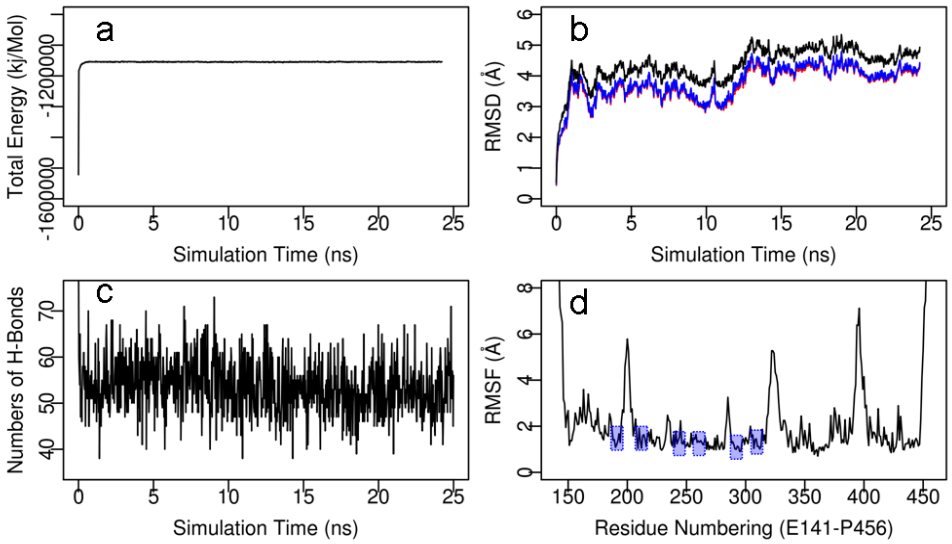


Figure 9. Analysis of a 25 ns MD simulation of the IRAK-M kinase domain. Fig. 9a, total energy was monitored during the simulation. Fig. 9b, root mean square deviation (RMSD) of the alpha carbon (Ca) (Blue), backbone (orange) and heavy atom (black) were calculated during the simulation. As the RMSD values of the backbone atoms are nearly identical to the RMSD values of the Ca atoms, therefore the orange curve cannot be observed clearly. Fig. 9c, the total number of hydrogen bonds in the domain along the MD simulation. Fig. 9d, the backbone RMS fluctuation of each residue in the domain was calculated, the x-axis represents the residue number from E141 to P456. The residues K192, E212, Y241, P242, Y243, R252, A297 and D311 positioned near the ATP binding pocket are indicated by blue rectangles.

During the MD simulation, the total energy of the kinase domain stabilized at -1.15 * 10^6 kJ/mol while the conformation of the model changed during the simulation until ~15 ns (Figure 9b). The average number of hydrogen bonds was 53 in the trajectory, which contribute to the overall 3D structural integrity. The fluctuation of each residue during the MD simulation was calculated and the flexible regions occurred in loops, all of which were not in close proximity of the ATP pocket (Figure 9d). The human IRAK-M kinase (Figure 8a) is comprised of two subunits which were linked by a hinge region P242-

L249 (PYMRNGTL), where N247-G248 is capable of forming an anti-parallel sheet with the strand I299-D302 (ILLD). The so-called gatekeeper residue Y243 of the kinase domain is located just N-terminal to the hinge region (Figure 8c). The kinase domain contained 9 alpha helix and 8 beta sheets. The proposed ATP-binding pocket (Figure 8d) had a volume of 624.5 $Å^3$, which is smaller than the ATP-binding pocket in IRAK-4 (752.1 $Å^3$). The non-spherical ratio of the proposed pocket was 1.56, which indicates that the pocket is narrow and deep (Figure 8d). The pocket was surrounded by residues K192, E212, Y241, P242, Y243, R252, A297 and D311. These residues were similar to their homologous residues in IRAK-4 except for V263 of IRAK-4, which had changed into P242 in IRAK-M (Figure 2b). Residue Y243 formed three hydrogen bonds with its neighbors A190 and E212. The kinase domain contains a unique loop M315- Y340, which is not found in other IRAK members. Residues T234 and H323 were solvent exposed in loop regions and K336 has three hydrogen bonds with G292, T355 and S359 (Figure 8e), which could explain the unusual orientation of these residues and thus why these residues are located in an unfavorable region of the Ramachandran plot (Figure 3).

The enzyme activity of a kinase involves the transfer of a phosphate group from ATP to a residue that contains a free hydroxyl group such as serine, threonine or tyrosine. Most kinases act on both serine and threonine residues, while others act on tyrosine only, or act on all three residues (dual-specificity kinases) (40). The kinase domain of human IRAK-4 was investigated from its crystal structure (41) and a number of important residues such as several glycine residues in the

P-loop, K213, Y264, D311 and T351 have been identified as being responsible for the kinase activity of human IRAK-4 (Figure 2a). In the majority of protein kinase structures, entry to the ATP binding pocket is blocked by amino acid side chains (for example F80 in CDK2, residues Q103/T106/M108 in most MAPK family members, or T315 in Abl kinase), and in IRAK-4, this so-called gatekeeper residue is a tyrosine (Y264). In the kinase domain of IRAK-M residue Y241 appears to control kinase activity by controlling the access to the ATP binding pocket, at a similar location as in its template IRAK-4. In the putative active site of hIRAK-M, S293 is present instead of a conserved aspartate that is essential for enzymatic activity (at position 311 in hIRAK-4). The hinge region of the kinase domain of IRAK-M connecting strand 5 and helix 3, which is the bridge between two subunits of the kinase domain, maintains the ATP binding pocket. Two loops in IRAK-4 have been suggested to be important for the expression of catalytic activity in IRAK family members (41): the activation loop at T351-Y354 (TTAY) and the phosphorylation loop (P-loop) at residue number G193-G198 (GEGGFG). However, when we study our IRAK-M model structure, although an activation loop and a P-loop are located at similar positions in the IRAK-M structure, these are not fully conserved and have changed into S332-Y340 (SSSSKHLWY) and G172-F177 (GEGEIF), respectively (Figure 2b).

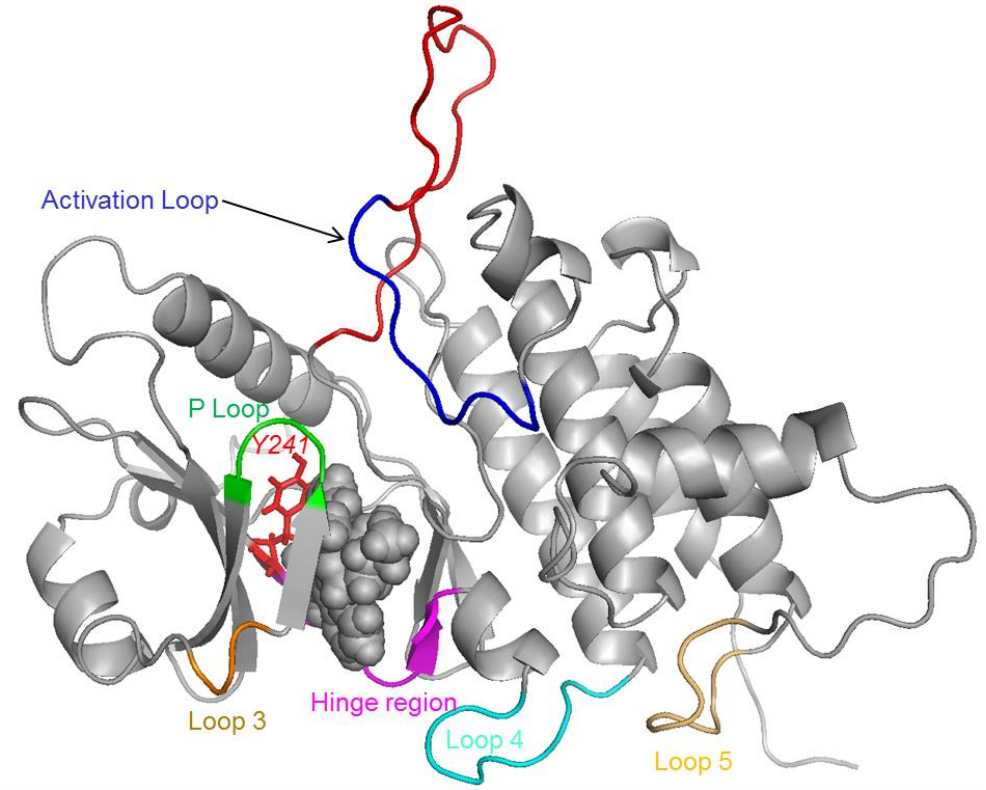


Figure 10. Loops in the IRAK-M kinase domain. The P loop (G172-F177) is shown in green; the activation loop (S332-Y340) in blue; the unique loop in red. The so-called loop 3 is shown in orange, loop 4 in cyan and loop 5 in light orange. The hinge region is shown in magenta. The proposed ATP binding pocket is filled with a pseudo substrate shown as a CPK representation. The side chain of gatekeeper residue Y241 is shown as stick type in red.

The presence of serine or threonine in the activation loop is required for active kinases. The presence of multiple serine residues (S333-S335) (Figure 2b) in the unique loop M315-Y340 in IRAK-M makes the geometry and polarization of the activation loop different from those in IRAK-4 because serine points outwards to the solvent and threonine points inwards to the core region of the kinase. The distance between loop3 (D168-L170) and loop4 (C255-P263, see Figure 10) is larger than that of IRAK-4, and therefore cannot

contribute to formation of catalytic activity. Loop 5 (Figure 10) which is in between two helices 5 and 6 of IRAK-M extends outward of the C-terminal lobe and points to the N-terminal lobe, forming a shield to the solvent.

Most kinases have a gatekeeper residue to control access to an internal ATP binding pocket (42). It is reported that nearly 20% of kinases in the human kinome possess the small residue threonine as a gatekeeper, 40% of them have methionine, of intermediate size, and nearly 15% have a large residue phenylalanine (43). However, a gatekeeper need not necessarily be threonine, methionine or phenylalanine, but it also can be one of the large hydrophobic residues such as isoleucine, valine, and tyrosine (43). The gatekeeper in IRAK-M is tyrosine (Y241). Around Y241, a big hydrophobic sphere is formed by F194, F209, L213, V215, L216, F312, A313. The distance between residue C238 and I299 represents the diameter of the pocket, 18.4 Å and a direct contact between these residues is not possible. However, these two residues still might be important for the stabilization of the ATP binding pocket, because Y241 is in between them. The distance from Y241 (Y262 in IRAK-4) to the ATP-binding site (K192), is at least 9.7 Å, whereas the distance in IRAK-4 is 9.0 Å.

Several other notable structural features are present in IRAK family members: A DFG motif is conserved amongst IRAK family members that are responsible for the coordination of Ca^{2+} or Mg^{2+} (43) and the third residue C-terminal of the DFG motif usually is an arginine (e.g. R334 in IRAK-4) (Figure 2b). The residue C-terminally to the primary

phosphorylation site (e.g. T345 and S346 in IRAK-4) is an arginine (e.g R347 in IRAK-4) (Figure 2b). The residue after the critical threonine (T352 in IRAK-4) in the P+1 pocket is tyrosine or histidine (41) (Figure 2b). In IRAK-M, however, the DFG motif of the active segment has changed into DFA (D311-F312-A313), and the third residue downstream of the motif is a histidine (H316) instead of arginine. The activation segment of IRAK-M does not have a serine phosphorylation site and next an asparagine is present instead of arginine (Figure 2b). The P+1 pocket shifts out of the center of the groove in IRAK-M because four consecutive serine residues are inserted here at the end of the P+1 pocket, even though the last residue is still tyrosine, like is the case in IRAK-4 (Figure 2b). For comparison, in the P+1 pocket of IRAK-4, there are many residues that contribute to substrate binding, such as R310, D311 and R334, by promotion of the correct orientation and electrostatic environment, or residues for the catalytic site, such as K313, or conserved threonine (T351) in kinase enzyme (Figure 2b). However, none of those residues is found in IRAK-M, which indicates that IRAK-M is an inactive kinase (Figure 2b).

Even though the architecture of the kinase domain of IRAK-M appears incompatible with the presence of kinase activity, it still is plausible that the IRAK-M kinase domain is involved in the down-regulation of NFκB dependent signaling since a region that is opposed to the ATP binding pocket was predicted as a potential interaction patch by three independent and basically different PPI programs (ODA, Cons-PPISP and PPI-Pred). The consensus ranked scores indicate a potential importance of residues L210, E214, L217,

M314, H316, C326, T327 for interaction with IRAK-M ligands or modulators. Thus, through means of a competition mechanism, IRAK-M may be able to bind to IRAK-ligands, thereby influencing binding equilibria.

The predicted residues have potential to interact with other proteins. Mutations on those residues may interfere with the IRAK-M functions. In order to conduct a rational mutagenesis experiment, we predicted the optimal residues to be mutated to and mutation prediction (38) suggested that the guided mutations L210K, E214S, L217N, M314D, H316F, C326N, T327S have potential to alter the IRAK-M interaction but not damage the structural integrity.

Conclusion

A model has been built for the death domain and kinase domain of human IRAK-M, a member of the IRAK protein family involved in toll-like receptor signaling. The models for both the hIRAK-M death domain and kinase domain have been analyzed by a consensus approach of several structural bioinformatics techniques in order to predict the most likely interaction areas that are involved in IRAK-M ligand binding and we have identified several areas on IRAK-M that we propose to be involved in protein-protein interaction. Several features of our models support the available functional data on IRAK-M. The absence of a typical active site and changes in the substrate binding pocket of the kinase domain of IRAK-M are in agreement with reports in literature that fail to associate this protein to kinase activity. At the center of the ATP-binding site, Y241, a so-called

gatekeeper may control access of small compounds to the hydrophobic pocket. Y241, as being exclusive gatekeeper to the IRAK family of kinases, interacts with E212 and then disturbs the salt bridge of E212 with K192. Combined with residue fluctuations as ascertained by MD simulation, hydrogen bond network analysis and homologue analysis, we propose mutagenesis of selected residues such as to alter the IRAK-M protein function while keeping the overall protein structure intact. These models are currently being used for the rational design of structure function studies through targeted mutagenesis of this important protein.

Acknowledgements

This work was supported by grants from the Transnational University Limburg (to GN), and by a grant from the China Scholarship Council to Jiangfeng Du (grant no. 2008630114).

References

1. Cao Z, Henzel WJ, & Gao X (1996) IRAK: a kinase associated with the interleukin-1 receptor. *Science* 271(5252):1128-1131.
2. Wesche H*, et al.* (1999) IRAK-M is a novel member of the Pelle/interleukin-1 receptor-associated kinase (IRAK) family. *The Journal of biological chemistry* 274(27):19403-19410.
3. Janssens S & Beyaert R (2003) Functional diversity and regulation of different interleukin-1 receptor-associated kinase (IRAK) family members. *Molecular cell* 11(2):293-302.
4. Lin SC, Lo YC, & Wu H (2010) Helical assembly in the MyD88-IRAK4-IRAK2 complex in TLR/IL-1R signalling. *Nature* 465(7300):885-890.
5. Suzuki N, Suzuki S, & Yeh WC (2002) IRAK-4 as the central TIR signaling mediator in innate immunity. *Trends in immunology* 23(10):503-506.

6. Maschera B, Ray K, Burns K, & Volpe F (1999) Overexpression of an enzymically inactive interleukin-1-receptor-associated kinase activates nuclear factor-kappaB. *The Biochemical journal* 339 (Pt 2):227-231.
7. Zhou H*, et al.* (2013) IRAK-M mediates Toll-like receptor/IL-1R-induced NFkappaB activation and cytokine production. *The EMBO journal* 32(4):583-596.
8. Li S, Strelow A, Fontana EJ, & Wesche H (2002) IRAK-4: a novel member of the IRAK family with the properties of an IRAK-kinase. *Proceedings of the National Academy of Sciences of the United States of America* 99(8):5567-5572.
9. Muzio M, Ni J, Feng P, & Dixit VM (1997) IRAK (Pelle) family member IRAK-2 and MyD88 as proximal mediators of IL-1 signaling. *Science* 278(5343):1612-1615.
10. Li X, Commane M, Jiang Z, & Stark GR (2001) IL-1-induced NFkappa B and c-Jun N-terminal kinase (JNK) activation diverge at IL-1 receptor-associated kinase (IRAK). *Proceedings of the National Academy of Sciences of the United States of America* 98(8):4461-4465.
11. Matthews RJ, Bowne DB, Flores E, & Thomas ML (1992) Characterization of hematopoietic intracellular protein tyrosine phosphatases: description of a phosphatase containing an SH2 domain and another enriched in proline-, glutamic acid-, serine-, and threonine-rich sequences. *Molecular and cellular biology* 12(5):2396-2405.
12. Darnay BG, Ni J, Moore PA, & Aggarwal BB (1999) Activation of NF-kappaB by RANK requires tumor necrosis factor receptor-associated factor (TRAF) 6 and NF-kappaB-inducing kinase. Identification of a novel TRAF6 interaction motif. *The Journal of biological chemistry* 274(12):7724-7731.
13. Ye H*, et al.* (2002) Distinct molecular mechanism for initiating TRAF6 signalling. *Nature* 418(6896):443-447.
14. Kobayashi K*, et al.* (2002) IRAK-M is a negative regulator of Toll-like receptor signaling. *Cell* 110(2):191-202.
15. Berman H, Henrick K, & Nakamura H (2003) Announcing the worldwide Protein Data Bank. *Nat Struct Biol* 10(12):980.
16. Neves MA, Totrov M, & Abagyan R (2012) Docking and scoring with ICM: the benchmarking results and strategies for improvement. *Journal of computer-aided molecular design* 26(6):675-686.
17. Berman HM (2008) The Protein Data Bank: a historical perspective. *Acta crystallographica. Section A, Foundations of crystallography*

64(Pt 1):88-95.
18. Joosten RP, Joosten K, Murshudov GN, & Perrakis A (2012) PDB_REDO: constructive validation, more than just looking for errors. *Acta Crystallogr D Biol Crystallogr* 68(Pt 4):484-496.
19. Joosten K, *et al.* (2008) A knowledge-driven approach for crystallographic protein model completion. *Acta crystallographica. Section D, Biological crystallography* 64(Pt 4):416-424.
20. Krieger E, Koraimann G, & Vriend G (2002) Increasing the precision of comparative models with YASARA NOVA--a self-parameterizing force field. *Proteins* 47(3):393-402.
21. Vriend G (1990) WHAT IF: a molecular modeling and drug design program. *Journal of molecular graphics* 8(1):52-56, 29.
22. Laskowski RA (2001) PDBsum: summaries and analyses of PDB structures. *Nucleic acids research* 29(1):221-222.
23. Bowie JU, Luthy R, & Eisenberg D (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science* 253(5016):164-170.
24. Colovos C & Yeates TO (1993) Verification of protein structures: patterns of nonbonded atomic interactions. *Protein science : a publication of the Protein Society* 2(9):1511-1519.
25. Pontius J, Richelle J, & Wodak SJ (1996) Deviations from standard atomic volumes as a quality measure for protein crystal structures. *Journal of molecular biology* 264(1):121-136.
26. Humphrey W, Dalke A, & Schulten K (1996) VMD: visual molecular dynamics. *J Mol Graph* 14(1):33-38, 27-38.
27. Fernandez-Recio J, Totrov M, Skorodumov C, & Abagyan R (2005) Optimal docking area: a new method for predicting protein-protein interaction sites. *Proteins* 58(1):134-143.
28. Chen H & Zhou HX (2005) Prediction of interface residues in protein-protein complexes by a consensus neural network method: test against NMR data. *Proteins* 61(1):21-35.
29. Bradford JR & Westhead DR (2005) Improved prediction of protein-protein binding sites using a support vector machines approach. *Bioinformatics* 21(8):1487-1494.
30. Vriend G (1990) WHAT IF: a molecular modeling and drug design program. *Journal of molecular graphics* 8(1):52-56, 29.
31. Pontius J, Richelle J, & Wodak SJ (1996) Deviations from standard atomic volumes as a quality measure for protein crystal structures. *Journal of molecular biology* 264(1):121-136.
32. L. DW (2007) The PyMOL molecular graphics system. *USA: DeLano Scientific*.

33. Morozov AV & Kortemme T (2005) Potential functions for hydrogen bonds in protein structure prediction and design. *Advances in protein chemistry* 72:1-38.
34. Feinstein E, Kimchi A, Wallach D, Boldin M, & Varfolomeev E (1995) The death domain: a module shared by proteins with diverse cellular functions. *Trends in biochemical sciences* 20(9):342-344.
35. Szperka ME, Connor EE, Paape MJ, Williams JL, & Bannerman DD (2005) Characterization of bovine FAS-associated death domain gene. *Animal genetics* 36(1):63-66.
36. Mendoza-Barbera E*, et al.* (2009) Contribution of globular death domains and unstructured linkers to MyD88.IRAK-4 heterodimer formation: an explanation for the antagonistic activity of MyD88s. *Biochemical and biophysical research communications* 380(1):183-187.
37. Neumann D*, et al.* (2008) Threonine 66 in the death domain of IRAK-1 is critical for interaction with signaling molecules but is not a target site for autophosphorylation. *Journal of leukocyte biology* 84(3):807-813.
38. Chinea G, Padron G, Hooft RW, Sander C, & Vriend G (1995) The use of position-specific rotamers in model building by homology. *Proteins* 23(3):415-421.
39. Eisenberg D, Luthy R, & Bowie JU (1997) VERIFY3D: assessment of protein models with three-dimensional profiles. *Methods Enzymol* 277:396-404.
40. Dhanasekaran N & Premkumar Reddy E (1998) Signaling by dual specificity kinases. *Oncogene* 17(11 Reviews):1447-1455.
41. Wang Z*, et al.* (2006) Crystal structures of IRAK-4 kinase in complex with inhibitors: a serine/threonine kinase with tyrosine as a gatekeeper. *Structure* 14(12):1835-1844.
42. Noble ME, Endicott JA, & Johnson LN (2004) Protein kinase inhibitors: insights into drug design from structure. *Science* 303(5665):1800-1805.
43. Alaimo PJ, Knight ZA, & Shokat KM (2005) Targeting the gatekeeper residue in phosphoinositide 3-kinases. *Bioorganic & medicinal chemistry* 13(8):2825-2836.

Chapter 6

General Discussion

In-silico experimentation in current biomedical sciences

With the development of computer sciences and of advanced algorithms for biophysical modeling, and born from a need to handle the ever-increasing amount of genetic and biological data that are being produced, application of *in-silico* approaches in the biomedical arena have become indispensable (1, 2). These technologies allow the construction of large data collections with well-known examples of such “big data” being the Protein Data Bank (PDB), the Zinc database, BindingMOAD, and even plausible biological resources like gene ontology (GO) and Online Mendelian Inheritance in Man (OMIM), which have established themselves as being fundamental to 21st century biomedical research (3). Programs such as FAFDrugs, ALOGPS, ToxPredict help researchers to filter or optimize large compound libraries prior to a HTS experiment by absorption, distribution, metabolism, excretion and toxicity properties (ADMEt) (4), thereby preventing the purchase and functional testing of non-drug-like compounds. While the availability of useful programs and software is not regulated and no true benchmarks appear to exist yet, initiatives by government funded agencies like the US-based National Institutes of Health (NIH) or the European Bioinformatics Institute (EBI) strive for harmonization and validation of methods. 3D structure models can assist experimentalists to understand the structure of target proteins such as to effectively screen compound databases (5). Knowledge of homology modeling may help experimentalists who employ NMR spectroscopy, x-ray crystallography or electron microscopy to elucidate 3D structures. In fact, modelling programs are an integrated part of the experimental

structure determination techniques and protein modeling has been developed to build 3D atomic structures within experimentally determined X-ray crystallographic density maps (6). Model building also provides the structural rational to assist experimentalists to perform mutagenesis studies, as a part of structure-functional analysis study (7, 8). Virtual ligand screening (VLS) methods, methods to discover new molecules with activities against a defined target, are widely used in the primary stages of drug discovery campaigns (9), and inclusion of VLS may result in considerable R&D budget reduction. Analysis of quantitative structure activity relationship (QSAR model) is being broadly applied for lead optimization (10-12) and is indispensable to modern drug design studies. Other structural bioinformatics techniques such as molecular dynamic simulation has the capacity to describe atomic properties such as protein structure formation, the binding process between drugs and their targets, protein conformational changes or cross-membrane transport , which are able to accelerate drug discovery, to provide the generation of new hypotheses or support the unraveling of mechanisms involved in biological processes (13-17).

The gap between *in-silico* and experimental results

Even though *in-silico* approaches have been booming in 21st century and are being routinely used by biomedical experimentalists, the methods are still under-exploited, mainly because of unawareness and because *in silico* techniques still are greeted with caution by some experimentalists (18). Bridging the gap between *in-silico* data and experimental results is not easy, unless both qualitative and

quantitative biological information is available and provided sufficient computational power is available. More so, information flow needs to be translated into human understandable language and there appears to be a language border between bioinformaticians and biomedical experimentalists, with the former tending to have a more biophysically/mathematically oriented and theoretical focus, while the latter have a biological/medical and much more pragmatic focus. For instance, for an experimental researcher or a medical doctor, an interesting value that refers to binding affinities should be a Kd value rather than a calculated binding energy, the latter seems to be too sophisticated and, while being scientifically sound and correct, is an over-abstraction of the description of a binding process and cannot be directly compared to experimental results. However, in molecular dynamics simulations, the binding energy is commonly used to interpret binding strength. Moreover, the gap between an experimental system and an *in-silico* system cannot be ignored. In an *in-vitro* experiment, a protein-protein interaction process may occur at a specific temperature, pressure, pH value, ionic strength, in buffer, in the presence of carrier protein and always in the presence of gravity. However, in an *in-silico* system, it is not possible to mimic all possible experimental conditions because of the limitation of computational power, or because of shortcomings in the technology itself. This consequently influences the difference between an MD simulation result and an experimental result. Limitations both from algorithms and applications exist in each *in-silico* approach, and a model that perfectly represents an experimental processes is currently not available, which in itself turn will maintain the gap

between *in-silico* results and experimental outputs (19). The difficulty of data integration from different data resources adds to the complexity of *in silico* experimentation, even though data combination is crucial for successful implementation of *in silico* methods in biomedicine and in drug discovery and design campaigns (20). Realization that there still exists a gap between *in silico* and *in vitro/in vivo* experimentation is crucial, and clear indication of limitations of technologies and interpretations, from both sides, in combination with translation of bioinformatics/biophysical data into human-understandable language, is key to successful synergy between *in silico* methods and more experimental biomedical techniques.

Optimization of combined VLS approaches

We addressed the question of how information about multiple target conformations can be included into a hierarchical structure-based virtual ligand screening (SBVLS) approach in the chapter 2 of this thesis. To do so, we used a multi-step protocol introduced by Miteva et al. in 2005, which in the case of the present study combines FRED and Surflex. Starting from a small database of ~60,000 molecules, we performed SBVLS on 10 different targets for which 8 conformations were prior generated for each target. We evaluate the SBVLS results in terms of enrichments and use 7 different methods of consensus scoring to rank the compounds. In addition, we evaluated two alternatives for how to optimally combine FRED and Surflex docking in a hierarchical manner. As we know, assembling the ligand dataset for evaluation of a VLS performance is essential.

To avoid a ligand dataset selection bias, a Directory of Useful Decoys (DUD) is commonly utilized to evaluate the quality of a VLS approach. However, the drawback of using the DUD is that the size of the compound database is rather small, it is intended for training and validation purposes and cannot be compared to commonly used commercial databases such as the Chembridge database (900,000) or ZINC databases (~35 millions). For example, the number of verified decoys for targets *FXa*, *thrombin*, *TK* and *NA* is 5745, 2456, 891 and 1874 respectively. In Chapter 2, we evaluate our VLS approach via two distinct approaches in order to provide a compromise with respect to both the ligand dataset selection bias and the dataset size. In the first approach, regarding the assembly of the ligand dataset, we use active ligands from DUD. Unconfirmed negative controls were retrieved from the ChemBridge database using random selection (~10% of the total compounds). This resulted a dataset with ~60,000 compounds. The dataset bias in this approach may exist if compared with 'standard decoys' from DUD. However, in this test, the goal is to find if the combined use of multiple conformers of a target perform better than single target structures when tested with the same dataset. So, in this respect, the dataset is unbiased to this aim. A drawback of the use of this approach however is that a quantitative performance of each approach cannot be identified. To this end, we used a second approach, in which we used the DUD dataset for the preparation of the active and inactive compounds, this set then can be used to calibrate the performance of our approaches and to compare our results with other VLS methods.

Amongst 7 consensus scoring methods, S_{opt} score was found to generally perform best. S_{opt} is a linear combination of the rank values of compounds with respect to the 8 conformations per protein target. However, in order to obtain S_{opt}, 9 parameters have to be fitted which makes the application of S_{opt} practically complex. Therefore, use of S_{opt} can only advised for well described targets and docking protocols and a simplified scoring function should be derived from S_{opt} method for more commonly encountered consensus scoring.

Keeping or removing water molecules in a binding pocket of a protein target is presently a much debated topic in HTS screening (21). It has been demonstrated that water molecules are important for molecular recognition (21). Moreover, it has been shown that water molecules in binding pockets influence the binding process significantly but that the effect of water on the docking process is not easily predicted, especially in HTS screening. In other words, the water molecule is important for target-ligand recognition, but inclusion of water in a binding pocket can make a HTS screening either better or worse. To this end, we consistently removed all the waters from the protein targets. During the development of scoring functions for water in VLS programs, such as GOLD, AUTODOCK-vina, water can be considered as a cofactor for VLS.

MD simulation of FVIIIa C domains

Activated biomembranes are of crucial importance for blood coagulation (22). For example, activation of the central enzyme thrombin by the prothrombinase complex (composed of FXa, FVa and membranes) is roughly 150,000 fold enhanced as compared to

activation by the complex of FXa, FVa in the absence of a membrane. In the homologous intrinsic tenase complex, the function of FIXa in the activation of FX and supported by the cofactor protein FVIIIa can be accelerated by >200,000 fold in the presence of phosphatidylserine containing membranes (23). Many studies have focused on the membrane binding domains of coagulation factors, either by experimental determination of protein 3D structures (24), by homology modeling and protein-protein docking (25), mutagenesis studies (26, 27) or by directly binding affinity measurements (28). Molecular dynamics (MD) is able to monitor the velocity of atoms, their interactions and energies in time. Therefore, provided there is an accurate force field, MD may provide accurate time-dependent data of atomistic detail for binding processes on a nanosecond or picoseconds time scale, which are difficult to be obtained via experiments methodologies. However a limitation of MD simulations is their time scale, currently nanoseconds level are mostly used (29). Many biological processes however require longer time scale. For example, protein folding requires hundreds of nanoseconds to minutes to complete, depending amongst others on protein size (29, 30). To expand simulation time scales, coarse-grained MD simulations have become attractive approaches to observe complicated biomolecular processes (31, 32), such as the interactions between the FVIII C domains and lipid membranes as described in chapter 3 of this thesis. In the chapter 3, by using the Martini coarse-grained MD simulation approach (33), we were able to observe how the FVIII C domains bind to a membrane, to describe which residues are important for membrane binding and investigate

the differences between wild type C domains and mutated variants thereof. Approximation of a biomolecular system by coarse-graining has its limitations. Conformational changes in a protein domain can be important during biological processes such as protein binding recognition, transportation and assembly (34-36). However, conformational changes cannot be observed in coarse-grained MD simulations because elastic network (EN) (37) are being applied in coarse-grained models to maintain the overall structure integrity, which inevitably restricts the conformational changes of a single domain model. In the multiple domain system (C1+C2), we removed all elastic network bonds between atoms which are in different domains (38), which however made it possible to measure the interdomain changes occurring between the C1 and C2 domains.

Since coarse-grained simulation greatly increases the time scale by simplification of a system, it reaches a less accurate description of free energy as compared to an MD simulation of atomistic detail. (31). The membrane binding energies calculated in the chapter 3 cannot be converted directly into binding affinity (Kd) since the binding energies of the C domains were notoriously low. The energies calculated in this chapter are however as such already suitable for a parallel comparison of the C domains and C-domain variants for which binding energies were derived by application of the same approach. A way to convert the current binding energies for the C-domain binding processes into experimental binding affinities has to be optimized.

It is reported that the FVIII A3 domain may contribute the FVIII

membrane binding process as well as C domains (25). We did however not include the A3 domain in our simulations and interpretation of our data does not allow us to comment on these earlier experimental studies and also was not the scope of our present study in which we specifically wished to address the binding of the FVIII C-domains to a membrane. From our results, we could see a synergy between the C1 and C2 domain upon membrane binding, which render a further question: whether there is a synergy between A3, C1 and C2 domains for the membrane binding? If the computational capability is allowed, a future direction should be the simulation of A3+C1+C2 membrane binding process or even the intact FVIIIa membrane binding process following a similar approach as was taken by us in the present case for only the binding of the C-domains.

IRAK-M project

In the chapters 4 and 5 of this thesis, we propose the location of interaction surfaces on the inflammatory adapter/mediator protein IRAK-M. This prediction was based on a homology modeling study using mouse IRAK-4 as a structural template. The proposed IRAK-M model is the basis for a series of site-directed mutants that are used to investigate the structure/function relationships of the IRAK-M death domain by stable expression in several cell types. The expression levels of wild type IRAK-M and those of the mutants have been measured after transfection of expression vectors for IRAK-M in HEK cells, so as to get an indication of overall correct folding. Meanwhile, we analyzed protein stability of a small selection of IRAK-

M death domain variants by MD simulation. Though our current results are sufficient to address our scientific goals, further MD simulations of several key mutants (such as F18, D19-A23 mutant, W74, R97 mutants) would be required to explore the conformational stability of these variant domains. In particular in case of the D19-A23 mutant, such information may support the interpretation of its reduced expression. From our mutagenesis results we observed that several mutants had differential effects on the capacity of IRAK-M to inhibit cytokine/chemokine production, dependent on cell type and stimulus. The complex results make the data interpretation difficult, which is why include “mild”, “moderately” and “major” terms to address the effects of IRAK-M mutants on NF-κB activity. In its turn this may increase the difficulty to understand these observations. Overall, we have studied the structure-function relationship of human IRAK-M death domain and identified the binding patches and the important residues for NF-κB signaling. Our results provided a structural context that helps in the understanding of the complex structure-function relationships of this protein that is involved in both CVD and in immune responses and the generated structures may contribute to future development of small drug-like molecules to interfere with the function of IRAK-M in NF-κB signaling.

1. Kapetanovic IM (2008) Computer-aided drug discovery and development (CADDD): in silico-chemico-biological approach. *Chemico-biological interactions* 171(2):165-176.
2. Gershell LJ & Atkins JH (2003) A brief history of novel drug discovery technologies. *Nature reviews. Drug discovery* 2(4):321-327.
3. Loging W, Harland L, & Williams-Jones B (2007) High-throughput electronic biology: mining information for drug discovery. *Nature reviews. Drug discovery* 6(3):220-230.
4. van de Waterbeemd H & Gifford E (2003) ADMET in silico modelling: towards prediction paradise? *Nature reviews. Drug discovery* 2(3):192-204.

5. Evers A & Klebe G (2004) Successful virtual screening for a submicromolar antagonist of the neurokinin-1 receptor based on a ligand-supported homology model. *Journal of medicinal chemistry* 47(22):5381-5392.
6. Langer G, Cohen SX, Lamzin VS, & Perrakis A (2008) Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. *Nature protocols* 3(7):1171-1179.
7. Mishra SK, Adam J, Wimmerova M, & Koca J (2012) In silico mutagenesis and docking study of Ralstonia solanacearum RSL lectin: performance of docking software to predict saccharide binding. *Journal of chemical information and modeling* 52(5):1250-1261.
8. Bromberg Y & Rost B (2008) Comprehensive in silico mutagenesis highlights functionally important residues in proteins. *Bioinformatics* 24(16):i207-212.
9. McInnes C (2007) Virtual screening strategies in drug discovery. *Current opinion in chemical biology* 11(5):494-502.
10. Verma J, Khedkar VM, & Coutinho EC (2010) 3D-QSAR in drug design--a review. *Current topics in medicinal chemistry* 10(1):95-115.
11. Ebalunode JO, Zheng W, & Tropsha A (2011) Application of QSAR and shape pharmacophore modeling approaches for targeted chemical library design. *Methods Mol Biol* 685:111-133.
12. Cumming JG, Davis AM, Muresan S, Haeberlein M, & Chen H (2013) Chemical predictive modelling to improve compound quality. *Nature reviews. Drug discovery* 12(12):948-962.
13. Ulmschneider MB, *et al.* (2013) Molecular dynamics of ion transport through the open conformation of a bacterial voltage-gated sodium channel. *Proceedings of the National Academy of Sciences of the United States of America* 110(16):6364-6369.
14. Gumbart J, Wang Y, Aksimentiev A, Tajkhorshid E, & Schulten K (2005) Molecular dynamics simulations of proteins in lipid bilayers. *Current opinion in structural biology* 15(4):423-431.
15. Borhani DW & Shaw DE (2012) The future of molecular dynamics simulations in drug discovery. *Journal of computer-aided molecular design* 26(1):15-26.
16. Padhi AK, Jayaram B, & Gomes J (2013) Prediction of functional loss of human angiogenin mutants associated with ALS by molecular dynamics simulations. *Scientific reports* 3:1225.
17. Dror RO, Dirks RM, Grossman JP, Xu H, & Shaw DE (2012) Biomolecular simulation: a computational microscope for molecular biology. *Annu Rev Biophys* 41:429-452.
18. Di Ventura B, Lemerle C, Michalodimitrakis K, & Serrano L (2006) From in vivo to in silico biology and back. *Nature* 443(7111):527-533.
19. Ekins S, Mestres J, & Testa B (2007) In silico pharmacology for drug discovery: applications to targets and beyond. *British journal of pharmacology* 152(1):21-37.
20. Searls DB (2005) Data integration: challenges for drug discovery. *Nature reviews. Drug discovery* 4(1):45-58.
21. de Beer SB, Vermeulen NP, & Oostenbrink C (2010) The role of water molecules in computational drug design. *Current topics in medicinal chemistry* 10(1):55-66.
22. Lentz BR (2003) Exposure of platelet membrane phosphatidylserine regulates blood coagulation. *Progress in lipid research* 42(5):423-438.
23. van Dieijen G, Tans G, Rosing J, & Hemker HC (1981) The role of phospholipid and factor VIIIa in the activation of bovine factor X. *J Biol Chem* 256(7):3433-

3442.
24. Liu Z, *et al.* (2010) Trp2313-His2315 of factor VIII C2 domain is involved in membrane binding: structure of a complex between the C2 domain and an inhibitor of membrane binding. *J Biol Chem* 285(12):8824-8829.
25. Stoilova-McPhie S, Villoutreix BO, Mertens K, Kemball-Cook G, & Holzenburg A (2002) 3-Dimensional structure of membrane-bound coagulation factor VIII: modeling of the factor VIII heterodimer within a 3-dimensional density map derived by electron crystallography. *Blood* 99(4):1215-1223.
26. Nicolaes GA, Villoutreix BO, & Dahlback B (2000) Mutations in a potential phospholipid binding loop in the C2 domain of factor V affecting the assembly of the prothrombinase complex. *Blood coagulation & fibrinolysis : an international journal in haemostasis and thrombosis* 11(1):89-100.
27. Gilbert GE, Kaufman RJ, Arena AA, Miao H, & Pipe SW (2002) Four hydrophobic amino acids of the factor VIII C2 domain are constituents of both the membrane-binding and von Willebrand factor-binding motifs. *J Biol Chem* 277(8):6374-6381.
28. Meems H, Meijer AB, Cullinan DB, Mertens K, & Gilbert GE (2009) Factor VIII C1 domain residues Lys 2092 and Phe 2093 contribute to membrane binding and cofactor activity. *Blood* 114(18):3938-3946.
29. Karplus M & Kuriyan J (2005) Molecular dynamics and protein function. *Proceedings of the National Academy of Sciences of the United States of America* 102(19):6679-6685.
30. Daggett V & Fersht A (2003) The present view of the mechanism of protein folding. *Nature reviews. Molecular cell biology* 4(6):497-502.
31. Tozzini V (2005) Coarse-grained models for proteins. *Current opinion in structural biology* 15(2):144-150.
32. Riniker S, Allison JR, & van Gunsteren WF (2012) On developing coarse-grained models for biomolecular simulation: a review. *Physical chemistry chemical physics : PCCP* 14(36):12423-12430.
33. Marrink SJ, Risselada HJ, Yefimov S, Tieleman DP, & de Vries AH (2007) The MARTINI force field: coarse grained model for biomolecular simulations. *J Phys Chem B* 111(27):7812-7824.
34. He HW, Zhang J, Zhou HM, & Yan YB (2005) Conformational change in the C-terminal domain is responsible for the initiation of creatine kinase thermal aggregation. *Biophysical journal* 89(4):2650-2658.
35. Granier S, *et al.* (2007) Structure and conformational changes in the C-terminal domain of the beta2-adrenoceptor: insights from fluorescence resonance energy transfer studies. *J Biol Chem* 282(18):13895-13905.
36. Wu S, Lee CJ, & Pedersen LG (2009) Conformational change path between closed and open forms of C2 domain of coagulation factor V on a two-dimensional free-energy surface. *Physical review. E, Statistical, nonlinear, and soft matter physics* 79(4 Pt 1):041909.
37. X. Periole MC, S.J. Marrink, M. Ceruso (2009) Combining an elastic network with a coarse-grained molecular force field: structure, dynamics and intermolecular recognition. *J. Chem. Th. Comp.* (5):2531-2543.
38. Siuda I & Thogersen L (2013) Conformational flexibility of the leucine binding protein examined by protein domain coarse-grained molecular dynamics. *Journal of molecular modeling* 19(11):4931-4945.

Valorisation

Cardiovascular diseases (CVD) are one of the main causes of death worldwide and it is estimated that 23 million people will die in 2030 from the diseases if no major breakthroughs in the development of CVD related drug discovery and treatments are to be made in the near future. Wet lab experimentation provides a major arena to study the pathological mechanisms of these diseases and to further explore ways for treatments. However, several reasons exist that push scientists looking for alternative methodologies to overcome shortcomings which exist in the research in the wet lab. For example, the traditional way to search for drug candidates from chemical databases is time consuming; therefore a financial budget is usually extremely high to conduct such researches in the regular drug discovery track. In addition, current technologies in the wet lab cannot reflect all biological reaction types; for example biological phenomena happening within femtosecond time scales. In our research, we have applied a so-called dry lab methodology, also popular known as *in-silico* approaches to the cardiovascular disease related area.

As a bioinformatician, one has to well understand not only the algorithms of the *in-silico* programs but also how to answer biological puzzles by application of these programs and further to bring achievement to our society. In the dissertation, we have provided an optimized virtual ligand screen (VLS) protocol aiming to speed up the drug discovery process, which will provide useful to the pharmaceutical industry or anyone who works in the drug discovery

field.

Drug discovery is a complicated, time consuming, expensive and ineffective process. In the pharmaceutical industry, people keep seeking novel approaches to cope with the steady increase of the number of disease-related proteins. High throughput screening (HTS) has become a routine method to filter chemical compounds in many labs. However, with the increased availability of protein structures and synthetic small compound databases, HTS does not work effectively for screening huge compound libraries. Therefore, in order to deal with the increased sets of compounds, rational improvements are needed to find a way to save time and control costs. To this end, virtual ligand screening (VLS) has been developed, such as to narrow down a huge compound library into a more manageable one at an early stage of drug discovery.

VLS is designed to speed up the drug discovery process, but which may be most beneficial in those instances where currently no pharmaceutical intervention/treatment is at hand but where VLS may be potentially applied. Continuously new drugs are needed, not only in cases where no drugs are available currently, but also to provide better alternatives that may be tailored to the needs of specific patient (sub)populations, such as for specific age-groups or for patients of differing genetic backgrounds.

In the present global economy, people, organisms and foods are easily transported to any corner of the world, which increases the risk of spreading of infectious diseases such as SARS, influenza or

ebola. The prevalence of these diseases has the potential to trigger social disasters if the infectious diseases are out of control. Novel medications are a major way to prevent epidemics, but sometimes the conventional drug discovery lags behind the spreading of diseases, such as the breakout of severe acute respiratory syndrome (SARS) in 2002 or the current outbreak of ebola in Western Africa. The application of VLS which accelerates the process of drug discovery has the potential to support the measures that are taken to prevent large-scale spreading of these diseases, by providing an expedited means of producing novel or improved drugs to treat the diseases, provided that enough information is at hand to perform this type of rationalized drug discovery and design.

Considering the diversity in protein structures, it is hardly possible to provide a generic and routine method that performs optimal for every protein target for a VLS campaign. Thus, numerous programs have been developed to execute a VLS campaign with respect to various requirements. In this respect, the combination of several different VLS programs more and more becomes common practice and a novel method called ‘stepwise improvement’, which is presented in our research can be used to improve speed and accuracy for VLS in drug discovery. The optimized protocols and the main suggestions made by us should be applicable to many flexible protein targets, but especially for the targets tested in the work: viz. thymidine kinase, neuraminidase, coagulation factor Xa, thrombin, glycinamide ribonucleotide transformylase, Cyclin dependent kinase 2, Catechol O-methyltransferase, Estrogen receptor and Enoyl ACP reductase.

Haemophilia A is a genetic disease caused by a deficiency of coagulation factor VIII (FVIII). The incidence of haemophilia A is nearly 1 in 10,000 males and the patients may suffer from bleeding, bruising and haematoma. A conventional treatment of haemophilia A is to replenish FVIII protein and to adjust FVIII plasma concentrations such that patients no longer bleed. However, immune resistance against FVIII is a major complication, which requires other solutions to manage the haemophilia A disease. In order to understand the function of Factor VIII and haemophilia A development, the mechanism of FVIII cofactor expression and regulation during blood coagulation has been well studied. We have unraveled the mechanism of the membrane binding of FVIII, which is a crucial step in coagulation pathway. The research has potential impact on haemophilia A drug discovery as it may allow development of small compounds to alter and improve Factor VIII membrane binding affinity. Likewise, our findings may proof useful for the development of novel antithrombotics, which may act through the inhibition of membrane binding of Factor VIII. In our research, we have predicted a list of membrane-binding residues, which may guide a gene therapy clinical trial designed for haemophilia A.

Our research on IRAK-M could have an impact on treatment of pneumonia, which occurs in 15% of the patients in the intensive care unit (ICU), of which 50% of the patients with pneumonia die. However, currently little progress has been made in this field and improved therapies are needed to overcome this problem. It is reported that IRAK-M is a down-regulator in immunoparalysis, which causes the secondary nosocomial pneumonia. Our research

provides a 3D map that may be used to rationally inhibit the activity of IRAK-M, which may potentially provide a long sought effective therapy for patients with temporary but life threatening immunoparalysis in the ICU.

Summary

Over the last decades, the application of *in-silico* approaches in cardiovascular research has been increasing. In this thesis, we describe the development of an optimal protocol to perform a virtual ligand screening campaign. We also applied various *in-silico* approaches to guide and support the structure-function study of IRAK-M. Further, we have conducted coarse-grained molecular dynamic simulation to study and simulate the membrane-binding process of FVIIIa, the haemophilia A protein.

In **chapter 2** of this thesis, we have developed a new method to perform structure based VLS and validated our approach by a stochastic compound database. We combined two structure based VLS programs FRED and Surflex in a parallel computational environment, aiming to accelerate the docking time while insuring the enrichment of hit molecules. Multiple conformations for each of a total of 10 protein targets have been generated so as to sample the conformational space that is available to the target as a result of intrinsic protein flexibility. We found that the use of multiple conformers for a given target is preferable over that of the use of a single target conformation. It is advisable to apply a consensus of the FRED docking results for consecutive Surflex screening. We have presented a novel approach (stepwise improvement) that is useful to detect the optimal cut-off from a rigid docking, as is FRED, to a flexible docking, like Surflex. To generate an optimal consensus result from eight conformations, we tested 7 different methods and we present a new consensus method, which is named S_{opt}, which is

designed to produce consensus lists for a protein target with an optimal enrichment for both FRED and Surflex. As described in chapter 2, we have tested the VLS performance on several cardiovascular related protein targets, including *FXa*, *thrombin*, *tyrosine kinase and propose the optimal approach for a VLS campaign for those targets. Moreover, several other homologous targets, that are of relevance to the cardiovascular system, can now be more optimally studied, guided by the findings of our studies.* In **chapter 3** of this thesis, an *in-silico* study has been performed to understand the membrane-binding mechanism of human FVIII, and more in particular that of the binding of the FVIII C domains to membranes containing phosphatidylserine. We constructed coarse-grained models for the C1, C2 and C1+C2 domains and also for seven variant domains for which have been studied experimentally before. By performing MD simulations we were able to study membrane binding times (the time it takes for a domain to stably bind to the membrane), to identify the individual residues that are involved in membrane binding and which of these residues become buried in the membrane including mode of membrane-burial, the depth of buried residues, the tilting of the domains, the C1+C2 angle movement and the membrane-binding energies. We also studied the relationship between 1,2-Dioleoyl-sn-glycero-3-phosphoserine (DOPS) lipid content in the membrane and the membrane-binding. We found that the C1 and C2 domains show different membrane binding properties. Some C domain variants had abnormal membrane-binding modes compared to that of the wild type. The C1 domain membrane binding was two times slower than that of the C2

domain, while the C1+C2 domain bound similarly to the C2 domain. The C domain variants which were experimentally confirmed to possess a higher binding affinity we found to bind faster than wild type C2 domain for binding. We concluded that the membrane binding speed is dependent on electrostatic interactions and identified the important residues for membrane-binding time. The tilting of the C domains, after they have been bound to the membrane, is likewise important for membrane-binding. The simulation results showed that 5% of DOPS lipids in the membrane is optimal for membrane-binding. Through analysis of the compiled contributions of electrostatic and van der Waals interactions and the identified membrane-buried residues, we present a list of residues which we identify as key residues for FVIII membrane binding. The results help to explain the molecular causes for haemophilia A, in those instances where membrane-binding is compromised, and predict likely candidates for novel causative missense mutations in haemophilia A.

In **chapter 4** of this thesis, we describe the structure-function relationship of the death domain of IRAK-M by analysis of mutant variants of this protein which have been rationally designed after creation and analysis of a 3D model for the IRAK-M death domain. The toll-like receptor (TLR) signaling inhibitor IRAK-M is a major player involved in the proper functioning of monocytes and macrophages. We generated a high quality 3D structure model for the IRAK-M death domain and identified two most likely interaction areas that are involved in IRAK-M ligand/modulators binding. These areas were then targeted by means of structure-guided mutagenesis.

We provided firstly a list of residues and/or combinations of several residues that are most likely to be involved in the interaction of the IRAK-M death domain with other protein ligands in the TLR pathway. We then prepared a series of recombinant IRAK-M mutants according to the *in-silico* prediction in different cell lines allowing the study of the function of these IRAK-M variant molecules. Furthermore, we proposed a structure for a tetramer of the IRAK-M death domain based on the X-ray structure of IRAK-4/2 tetramers (3MOP.pdb). The models and mutagenesis results showed that the NF-κB activating activity of IRAK-M is dependent on two different sites on the DD, with respectively W74 and R97 as critical IRAK-4 binding residues. Residues W74 and R97 may interact with IRAK4 and prevent the formation of the IRAK-4-IRAK-2-MyD88 complex and further inhibit the TLR2 and TLR4 mediated TNF and IL-6 production in human monocytes. Our results suggest that the tetramers may sandwich an IRAK-4-DD tetramer between the R97-exposing surface and the W74-exposing surface that are situated at respectively the top and bottom of the IRAK-4 tetramer. Residues W74 and R97 are also important for NF-κB and ERK activation as well as for the inhibitory action of IRAK-M on TLR induced release of cytokines. Residue R70 and the stretch D19-A23 are specifically involved in ERK activation and IRAK-M expression levels. Our study provides novel insights in the molecular mechanisms of IRAK-M by differential use of its death domain. This may channel a way for the rational design of IRAK-M inhibitors, which may be applied in future treatment strategies and which could potentially improve innate immunity in vulnerable patients or CVD patients.

In the **chapter 5** of this thesis, we expanded the in-silico approaches from chapter 4 to the kinase domain of IRAK-M. In this chapter, we analyzed the IRAK-M protein solely via bioinformatics methods. By homology modeling and data analysis, we confirmed that the likely reason that causes the kinase domain to be inactive. For example, the kinase domain lacks a primary phosphorylation site, followed by a histidine instead of an arginine at the P+1 site. Serine residues are missing in the activation region and the ATP binding pocket is too narrow to accommodate ATP substrate although tyrosine (Y241) is present as the gatekeeper residue in the pocket. The chapter 5 provides a more extensive description of the 3D structure of the death domain as compared to chapter 4 and moreover describes the 3D structure of the kinase domain. The quality of the structures is discussed after evaluation by means of different evaluation factors such as 3D packing quality, bond quality, hydrogen bonding and structural stability. The IRAK-M kinase model structure is compared to that of other active kinase domains. Finally, residues which are most likely to interact with the kinase substrates are predicted.

Samenvatting

De toepassing van onderzoek dat expliciet gebruik maakt van computers, ook wel *in silico* methoden genoemd, heeft gedurende de laatste tien tot twintig jaar ook binnen het cardiovasculaire onderzoeksveld een vlucht genomen. In dit proefschrift wordt de ontwikkeling van een optimaal protocol beschreven dat gebruikt kan worden voor het vinden van kleine moleculen die binden aan een bepaald doeleiwit door middel van zogenaamde virtuele ligand screening. Verder hebben wij *in silico* methoden toegepast om de structuurfunctie studie van IRAK-M te sturen en te ondersteunen. Tenslotte hebben wij grofschalige moleculaire dynamica simulaties uitgevoerd, om daarmee de binding van FVIIIa, het hemofilie A eiwit, aan een lipide membraan te kunnen simuleren en te kunnen bestuderen.

Hoofstuk 2 beschrijft de ontwikkeling van een nieuwe methode om struktuurgedreven virtuele ligand screening te bedrijven. Hierbij zijn twee bestaande programma's, FRED en Surflex, in een parallelle rekenomgeving toegepast, met als doel om de totale tijd van de screening te versnellen en de doelmatigheid van de gehele procedure te verhogen. Hierbij zijn er tien model eiwitten gebruikt om onze nieuwe methode mee op te zetten. Door gebruik te maken van verschillende conformaties van eenzelfde doeleiwit verkregen wij betere resultaten dan wanneer wij slechts een enkele conformatie van dit eiwit gebruikten. Om te bepalen hoe het beste de beide programma's, in combinatie met het gebruik van meerdere doeleiwit structuren konden worden gecombineerd, hebben wij zeven

verschillende methodes getest. Eén van deze methoden, Sopt genaamd, presteerde beter dan de andere consensus methoden.

Hoofdstuk 3 van dit proefschrift beschrijft een *in silico* studie die is uitgevoerd om de membraan bindende eigenschappen van de humane stollingsfactor VIII te bestuderen, en meer in het bijzonder de binding van de carboxy-terminale C-domeinen van factor VIII aan modelmembranen die het fosfolipide molecuul fosfatidylserine bevatten. Door middel van grofschalige moleculaire dynamica experimenten waren we in staat de bindingstijden te berekenen die behoren bij de binding van een C1 domein, een C2 domein of een C1+C2 domein aan een lipidelaag. Hierbij waren we in staat individuele aminozuren te identificeren die betrokken zijn bij het bindingsmechanisme en meerdere individuele parameters te kwantiteren die de binding van C-domeinen aan de membraan beschrijven.

Naast de gesimuleerde wildtype domeinen, werden er ook enkele gesimuleerde varianten van de C-domeinen getest ter controle van de juistheid van de door ons gebruikte *in silico* methoden. In alle gevallen waren de gevonden resultaten in overeenstemming met data uit de literatuur. Dit hoofdstuk wordt afgesloten met een voorstel dat het bindingsmechanisme beschrijft waarmee C-domeinen op calcium onafhankelijke wijze kunnen binden aan negatief geladen membranen.

Met dit hoofdstuk presenteren wij niet alleen een algemeen mechanisme voor binding waarbij individuele aminozuren kunnen

worden aangewezen die van groot belang zijn voor de membraan binding van factor VIII. Deze informatie en de door ons toegepaste *in silico* simulaties kunnen gebruikt worden om een moleculaire verklaring te geven voor de bloedingsneiging die bij hemofilie A patiënten bestaat, zowel in geval van bekende mutaties, maar ook voor nieuw ontdekte mutaties.

Hoofdstuk 4 van dit proefschrift geeft een beschrijving van de structuurfunctie relaties van het zogenaamde death domein van humaan IRAK-M, een eiwit dat deel is van het menselijk immuunsysteem. De manier waarbij we hier onderzoek hebben verricht is door het maken van een 3D homologie model voor dit eiwit domein en dit model vervolgens te gebruiken om voorspellingen te doen over potentiële interactie gebieden tussen IRAK-M en zijn bindingspartners. Deze informatie is op zijn beurt gebruikt om op rationele wijze mutaties aan te brengen in het IRAK-M. Door functionele analyse te doen van, tot expressie gebrachte recombinante varianten van het IRAK-M, is er waardevolle informatie verkregen over de functie van dit eiwit. Belangrijke aminozuren zijn geïdentificeerd en voorts is er een voorstel gedaan over een mogelijke quaternaire structuur waarin het IRAK-M een remmende rol zou kunnen vervullen door complexering met IRAK-4.

Hoofdstuk 5 beschrijft een uitbreiding van de *in silico* studies die gedaan zijn aan IRAK-M door ook expliciet het kinase domein van humaan IRAK-M te bestuderen. Het betreft hier uitsluitend de toepassing van bioinformatica methodes waarbij een 3D model is gemaakt voor het kinase domein van IRAK-M. Bestudering van dit

model, en een vergelijk met andere kinase structuren geeft ons een beter inzicht in de functionaliteit van dit domein, en geeft met name een verklaring waarom dit kinase inactief is. Tenslotte beschrijft dit hoofdstuk een uitgebreide analyse van de 3D structuur van het IRAK-M death domein, die voortbouwt op de reeds in hoofdstuk 4 genoemde resultaten.

总结

在过去的几个十年中，生物信息学方法广泛的应用于心血管基础医学领域。在该部论文中，我们优化了计算机辅助药物筛选(virtual ligand screening, VLS)的算法；运用结构生物信息学和点突变来研究免疫蛋白 IRAK-M 的结构与功能；运用分子动力学计算模拟研究血友病 A 型相关凝血因子 FVIIIa 并阐释其与细胞膜结合的机理。

在本论文的第二章中，我们优化了基于结构 VLS 的算法，并通过随机选取小分子数据库来验证此算法。该算法加快了筛选时间，同时也提高了筛选准确率。本实验选取十个靶蛋白并且使每个蛋白分子产生八个构象以便于模拟蛋白质高级结构的可变性。我们发现，使用多种构象做虚拟筛选的效果优于仅仅使用单一构象。我们推荐使用整合后的 FRED 结果然后进而运行 Surflex 筛选。我们提出了一种新的方法（逐步提高检测法）对刚性筛选结果的选择是有效的。为了得倒最佳筛选效果，我们测试了七种不同的方法。其中最佳的组合方法被命名为 S_{opt}。正如第二章所述，我们已经在多个心血管疾病相关蛋白分子，例如凝血因子 Xa，凝血酶，酪氨酸激酶上面做了实验验证，并为这些蛋白靶点提供了最佳筛选参数。第 3 章描述了结构生物信息学在研究人类凝血因子 FVIII 的膜结合的生物学机制。我们构建了 FVIII 蛋白分子 C1，C2，C1 + C2 结构域和七个突变体的三维模型。通过进行分子动力学模拟，我们能够计算这些模型结合到细胞膜上所需的时间（结合时间），并且能够鉴别出直接作用于细胞膜结合的具体氨基酸残基，插入到细胞膜的深度，结构域的最佳结合角度以及结合能。我们还研究了在细胞膜中 1,2 - 二油酰基-sn-甘油基-3 - 磷酸丝氨酸

（DOPS）脂质含量和膜结合之间的关系。我们发现，C1 和 C2 结构域具有不同的膜结合特性。相比于野生型中的 C 结构域，有些 C 结构域的突变体有异常膜结合模式。C1 结构域的膜结合时间比的 C2 结构域慢 2 倍，而 C1 + C2 结构域的结合时间基本类似于 C2 结构域。实验验证具有比野生态结合能高的突变体也具有比野生态更短的结合时间。我们推论 C 结构域的膜结合特性依赖于静电相互作用，并鉴定出与结合时间相关的重要氨基酸残基。当结合到细胞膜后，C 结构域的倾斜被认为对其结合能力产生重要影响。研究结果表明，当细胞膜里面含有 5%的 DOPS，C 结构域的膜结合效果最佳。结合静电和范德华相互作用的因素，我们预测出一系列参与到细胞膜结合的氨基酸残基。该研究结果有助于从分子水平解释血友病 A 的致病原因。

在本论文的第四章中，我们通过分析所构建的 IRAK-M death domain 三维结构的模型，并结合点突变实验研究了其结构与功能的关系。Toll 样受体（TLR）信号传导抑制剂 IRAK-M 是参与单核细胞和巨噬细胞的正常运作的主要蛋白。我们构建了高品质的 3D 结构模型，并预测了两个配体结合区域。我们在这些区域研究其氨基酸定点突变。我们得出了一系列最有可能参与到分子互作的氨基酸。此外，基于模板晶体结构，我们构建了 IRAK-M death domain 的四聚体的结构。该模型和点突变研究结果表明，IRAK-M 在 NF-κB 激活途径依赖两个不同的位点，W74 和 R97 分别为关键氨基酸。氨基酸残基 W74 和 R97 可与 IRAK4 相互作用并阻止 IRAK-4-IRAK-2-MyD88 的复合物的形成，并进一步抑制 TLR2 和 TLR4 介导的 TNF 和 IL-6 的产生。研究结果表明该四聚体可能夹裹 IRAK-4 四聚体。在 NF-κB 和 ERK 的激活途经中，氨基酸 W74 和 R97 同样行使重要功能。R70 和 D19-A23 区域可能特异的

参与到 ERK 活化过程和影响到 IRAK-M 表达水平。本研究揭示了 IRAK-M 参与到多种生物途经的分子机制。本实验为针对 IRAK-M 为靶点的理性药物研发提供了重要思路。这可能被应用到未来治疗心血管方面的疾病或者改善患者的生活质量。

在本论文的第五章中，我们将生物信息学的方法从 death domain 拓展至激酶结构域。我们从生物信息学方面分析了 IRAK-M 激酶结构域没有活性的原因。例如，该激酶结构域缺乏主磷酸化位点，在随后的位点上组氨酸代替精氨酸。激活区域缺失丝氨酸残基，并且 ATP 结合区域太窄进而无法容纳 ATP 底物。本章详细的描述了 IRAK-M 的两个结构域的三维结构信息并讨论了其三维模型的质量，例如分析其原子排列质量，分子键，氢键和结构稳定性。研究对 IRAK-M 激酶 ATP 结合区和潜在的催化位点做了充分的描述，并对比了其他具有活性的激酶。最终，一些最有可能参与到其激酶底物结合的氨基酸被预测。

Publications

Du, Jiangfeng, Bleylevens, I. W. M., Bitorina, A. V, Wichapong, K., & Nicolaes, G. A. F. (2014). Optimization of compound ranking for structure-based virtual ligand screening using an established FRED-Surflex consensus approach. *Chemical biology & drug design. doi:10.1111/cbdd.12202*

Wu, Y., **Du, J**., Wang, X., Fang, X., Shan, W., & Liang, Z. (2012). Computational prediction and experimental verification of miRNAs in Panicum miliaceum L. *Science China. Life sciences*, *55*(9), 807–17. doi:10.1007/s11427-012-4367-y

Du, JiangFeng, Wu, Y., Fang, X., Cao, J., Zhao, L., & Tao, S. (2010). Prediction of sorghum miRNAs and their targets with computational methods. *Chinese Science Bulletin*, *55*(13), 1263–1270. doi:10.1007/s11434-010-0035-4

Du, JiangFeng, Wu, Y., Zhang, Y., Wu, L., Wang, X., & Tao, S. (2009). Large-scale information entropy analysis of important sites in mature and precursor miRNA sequences. *Science in China. Series C, Life sciences / Chinese Academy of Sciences*, *52*(8), 771–9. doi:10.1007/s11427-009-0099-z

Manuscripts in preparation

Jiangfeng Du, Gerry A.F. Nicolaes, Danielle Kruijswijk, Tom van der Poll, Miranda Versloot and Cornelis van ’t Veer. The structure function of the death domain of human IRAK-M. Submitted to Cell Communication and Signaling

Jiangfeng Du, Wichapong, K., Hackeng TM, Gerry A.F. Nicolaes. Molecular dynamic simulations of human coagulation factor VIII C domain-mediated membrane binding**.** Submitted to Thrombosis and Haemostasis

Jiangfeng Du, Cornelis van’t Veer, Gerry A.F. Nicolaes. Homology modeling and binding site prediction of human IRAK-M. Submitted to Chinese Science Bulletin

Selected Abstracts

Jiangfeng Du, Kanin Wichapong, Gerry A. F. Nicolaes. MD Simulation Studies of the membrane binding process of the Human Blood Coagulation Factor VIII C2 Domain. Oral presentation at the 3rd Dutch Molecular Dynamics (DMD), March 08, 2013, Eindhoven, the Netherlands.

Jiangfeng Du, Kanin Wichapong, Gerry A. F. Nicolaes. MD Simulation Studies of the membrane binding process of the Human Blood Coagulation Factor VIII C Domains. Main poster presentation at the XXIV International Society on Thrombosis and Haemostasis, June 29 – July 4, 2013, Amsterdam, the Netherlands.

Jiangfeng Du, Kanin Wichapong, Gerry A. F. Nicolaes. MD Simulation Studies of the membrane binding process of the Human Blood Coagulation Factor VIII C Domains. Oral presentation for Netherlands Society on Biomolecular Modeling (NSBM). October 25, 2013, Maastricht, the Netherlands.

Jiangfeng Du, Ivo W.M. Bleylevens, Albert Virgilio Bitorina, Gerry A. F. Nicolaes. Optimization of compound ranking for structure-based virtual ligand screening using an established FRED-Surflex consensus approach. Poster presentation for Netherlands Society on Biomolecular Modeling (NSBM). November 26, 2012, Utrecht, the Netherlands.

Curriculum Vitae

Jiangfeng Du was born on November 10th, 1982 in SanMenXia, Henan Province, People's Republic of China, where he obtained his education in a primary and high school. In September 2002, he attended to Northwest A&F University (Xi'an, China) in Biological Sciences as a bachelor student and later he obtained his bachelor's degree in Science and a secondary bachelor's degree in Computer Engineering. Afterwards, he continued his study in the field of bioinformatics and MicroRNA as a MSc student from September 2006 to July 2009 under the supervision of professor Shiheng Tao and associated professor Yongjun Wu in the same university. In July 2009, he graduated as a Master of Science in Bioinformatics. In December 2008, he received a Chinese Scholarship Council (CSC) scholarship from China government, which granted his 4-year PhD research from 2009 to 2013 in the Department of Biochemistry at the Cardiovascular Research Institute Maastricht (CARIM) of Maastricht University, the Netherlands. The PhD research was supervised by Dr. Gerry A.F Nicolaes and Prof. Tilman Hackeng. During his PhD period, he attends a number of symposia and congresses, which are listed below:

SYMPOSIUM

NSBM meeting: October 25, 2013 (oral presentation)

ISTH : *June 29 - July 4, 2013*

The International Society on Thrombosis and Haemostasis(Poster/oral presentation) (Amsterdam)

NBIC: *April16-17, 2013*

The Netherlands Bioinformatics Conference (poster) (Lunteren)

DMD: *March 8, 2013*

Dutch Molecular Simulation Day (Oral presentation) (Eindhoven)

LD symposium: *November 30, 2012*

Medicinal Chemistry of SRC & KVCV, MedChem, 2012 (Poster presentation) (Liège, Belgium)

NBIC*: April 16-17, 2012*

The Netherlands Bioinformatics Conference (poster) (Lunteren)

DMD *March 23 2012* (Poster) (Groningen)

ICCS: *June 5-9, 2011*

International Conference on Chemical Structures (Noordwijkerhout, the Netherlands)

NVTH: *May 17-18, 2011*

Nederlandse Vereniging voor Trombose en Hemostase (Abstract)

NBIC: *April 16-17, 2011*

LD symposium: *September 5-9, 2010*

21st International Symposium on Medicinal Chemistry (Brussels)

NVTH & BSHT June 23-25, 2010

British Society of Hemostase en Thromobosis Joint Symposium (Noordwijkerhout, the netherlands)

Acknowledgements

At the end of the thesis, I feel guilty if I skip expressing my appreciations to the people who have accompanied me on my way in the intensive, challenge and beautiful years in the Netherlands during my Ph.D career.

The first sincere gratitude is delightful to my co-promotor Dr. **Gerry A.F Nicolaes**, who is the most important person for my life in the Netherlands, and over the past few years he continuously gave me supports in both science and non-science matters. As a direct supervisor, he almost participated in all the important events I had in these years. He helped me improved my scientific skills, which paved a path to my scientific career. It is his connections in the academic world that facility my corporation between peers. Without his valuable comments and suggestions throughout the thesis, my thesis would be difficult to be presented here. I am very happy to have the opportunity to work in his group. I also love to thank his wife and the cute twins for their hospitalities in the past four Saint Nicholas events. Thanks again!

My promotor, Prof. Dr **Tilman Hackeng**, he is a very generous person that I have never met before. He kindly granted me another half year extension in the Netherlands for my PhD research. Without his support, it is not possible to finish the last part of the work. Many thanks are given to him for his time on the thesis. It is awesome that any big tasks became small when I came to him. I also want to say sorry that sometimes I was nervous (in Chinese-style) when I

discussed things with him because he is head of the department.

My special appreciation is given to Dr. **Cornelis van 't Veer**, Center for Experimental and Molecular Medicine, University of Amsterdam and **Ivo Bleylevens**, Department of Genetic, Maastricht University. It is my pleasure to work together with them. I learned a lot from the collaboration. I appreciate their efforts in the thesis.

My colleague **Karin**, she is so sweet and helpful. Recalling back in the past four years, I realized that we together attended countless social activities and PhD events. I would say that my life could be tough, difficult, boring, sorrow, cloudy, gloomy if she was not on beside me. She is really special to me. I would like to give my gratitude to **Simone**. From the first days I came into our department, she introduced me to our department. Who knows how fragile I was at that moment, standing behind her, seeking for protection. From then on, she always cared of me and she had a real talent to capture persons' feelings, at least to me. She was always on my side and fought for my right, maximizing my profits. Thanks a lot. I never called it a subtle event when she pushed me into Tilman's office, which was a big step to make my PhD completed. Thanks once again!

I like to give my acknowledgement to my previous co-worker, Dr. **Mahesh Kulharia**, a nice and wisdom Indian guy. I remember that in the first few days in the Netherlands he gave me hands in scientific research, as well as in other aspects about life, nature, literature, politics and English.

I am very happy to have **Barbara** and **Kanin** working together in the lab. We shared many ideas in energy calculation, molecular docking algorithms. There were many unforgettable time we spent together in the past few years. It was a lot of fun that I missed the last train with Kanin to Maastricht, which made us wandering in Utrecht until the first train in next day. I appreciated that Barbara guided me a trip in her motherland, Poland, where I got many cheerful memories. I wish her a big success. **Stella Thomassen**, she is not just a colleague of mine, but also a very nice friend. She is so kind to say “ni hao” that slightly erased my homesickness. She got talent in communication, and organizing things, which is something I have been learning from her. I am grateful that she encouraged me joining in various social activities such as Christmas running, Badminton events, drinking in bars, dancing in parties. **Dr. Elisabetta Castoldi**, also famous as Betta, do you remember that we cycled together to the border of the country? Do you remember we often saw each other in the late evening in the department? Thank you very much that you always came to me and comforted my soul when I was down. **Roy Schrijver**, did you notice that I like to talk with you? It was delightful to chat with him. We sometimes could talk in deep for a topic, which was awesome. I also appreciated that he offered me some furniture.

Lunch time was an important media to share news, spread “gossip” and enhance friendship, so I am delightful to give my appreciation to the people: **Alexandra, Brecht, Connie, Farida, Francesca, Hans, Ingrid, Kristien, Linda, Lisbeth, Marie, Mehash, Olivier, Pieter, Stijn, Sameera.** I want to express my appreciation to the respected professor **Dr. Jan Rosing**.

I express my gratitude to the people who helped me directly from CMBI, Nijmegen: **Prof. Vriend**, **Hanka, Barbara van Kampen, Martin**; from university of Groningen: Prof. **Siewert-Jan Marrink, Clement, Djurre, Floris, Lars, Xavier**. I was grateful that **Dr. Bruno Villoutreix** and **Sperandio Olivier** for their visiting to our group.

Many thanks were delivered to **Trees, Elsa** and **Lidewij**. Their professional support and assistance made me fully dedicated to the scientific research. I felt happy that **Trees** took care of me like my grandma in some degrees. I wish her a wonderful, healthy life!

My gratitude to the members of my dissertation committee **Prof. Dr. Erik Biessen, Prof. Dr. Jan Glatz, Prof. dr. Siewert-Jan Marrink,** and **Dr. Bruno Villoutreix** for their fast evaluation and positive responses. I appreciate the time they have spent to my dissertation.

The thesis is dedicated to my parents and my younger brother, whose unconditional spiritual and material supports from China made it possible to get so far. It is pity that none of them would be able to come to my Ph.D ceremony to share the exciting, joyful and meaningful moment with me. I gave my deep gratitude to **Q. Zhang**, who lent me encourage, willing. My life in the Netherlands was colorful and unforgettable in company with her.

In addition, I would like to say some words to the friends we had spent time together. Among them I would like to mention by name: **Junfang Zhao, Jieyi Li, Yimeng Song, Zhengqiu Chen, Zhengyang Lu**. I owe also a lot to my badminton mates, who have offered a perfect environment to smash and sweat.