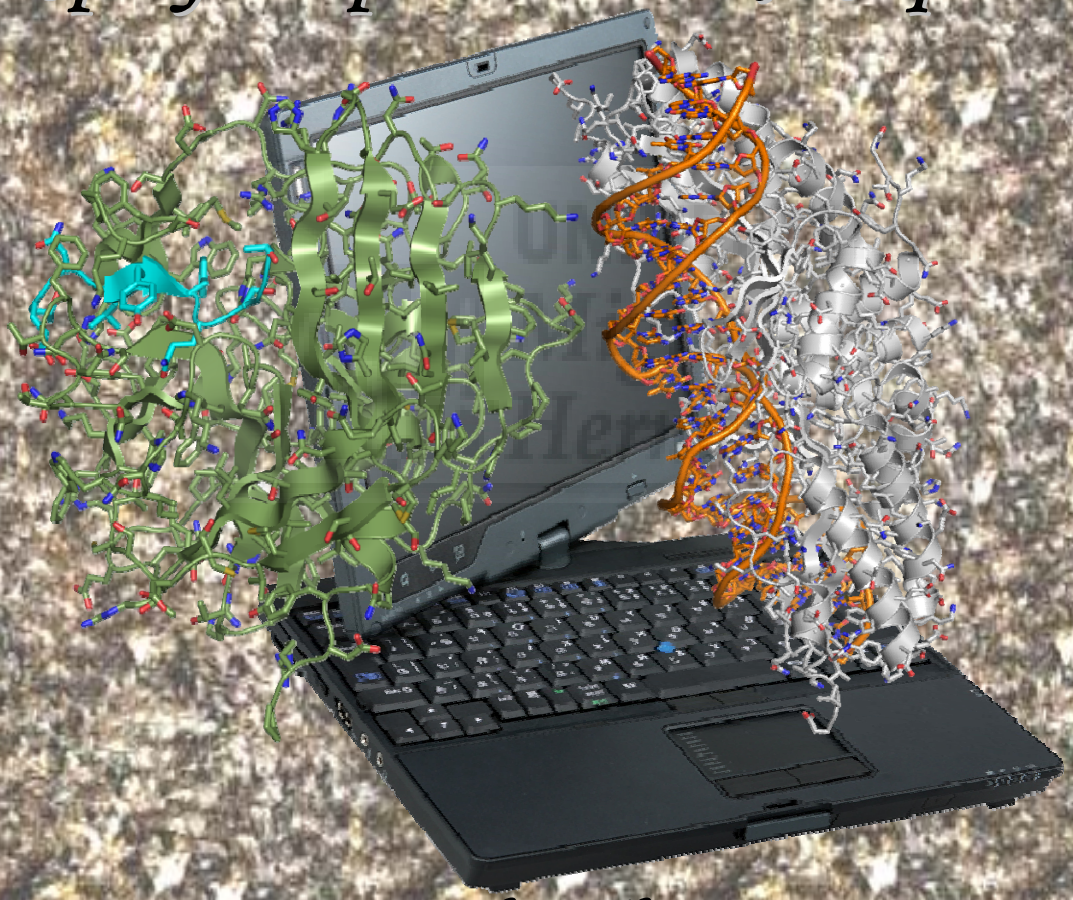
Universidad Miguel Hernández
Instituto de Biología Molecular y Celular

"Computer-Aided Protein Design for Engineering Enzymes to Recognize Specific Peptide and DNA Sequences"



PhD Thesis

Emmanuel Fajardo Sánchez

Elx, 2007

UNIVERSITAS
Miguel
Hernández

Universidad Miguel Hernández
Instituto de Biología Molecular y Celular

"Computer-Aided Protein Design for Engineering Enzymes to Recognize Specific Peptide and DNA Sequences"

PhD THESIS

EMMANUEL FAJARDO SÁNCHEZ

Elx, 2007

Thesis Directors

Prof. Dr. Luis Serrano Pubull (EMBL)

Prof. Dr. Gregorio Fernández Ballester (UMH-IBMC)

“Computer-Aided Protein Design for Engineering Enzymes to Recognize Specific Peptide and DNA Sequences”

PhD Student

EMMANUEL FAJARDO SÁNCHEZ

Born in Alicante, (Spain)

on the 12th of March 1977.

Thesis Directors

Prof. Dr. Gregorio Fernández Ballester (UMH-IBMC)

Prof. Dr. Luis Serrano Pubull (EMBL)

Instituto de Biología Molecular y Celular (IBMC) UNIVERSIDAD MIGUEL HERNANDEZ (UMH)

Campus de Elche. España

Structural and Computational Biology Unit EUROPEAN MOLECULAR BIOLOGY LABORATORY (EMBL)

Heidelberg. Germany

Prof. Dr. José Manuel González Ros, director of the Cellular and Molecular Biology Institute from the Miguel Hernández University of Elche (Spain),

GIVES THE APPROVAL for the defence of the PhD Thesis, titled: "Computer-Aided Protein Design for Engineering Enzymes to Recognize Specific Peptide and DNA Sequences", submitted by the Bachelor of Biology Mr. Emmanuel Fajardo Sánchez.

In witness whereof and for all pertinent purposes, the present certificate is hereby issued in Elche, on the 12th June 2007.

Signed: Prof. Dr. José Manuel González Ros

Dr. Gregorio Fernández Ballester, titular professor of Biochemistry, in the department of Molecular biology, within the Cellular and Molecular Biology Institute of the Miguel Hernández University of Elche (Spain),

and

Prof. Dr. Luis Serrano Pubull, coordinator of the Structural and Computational Biology Unit of the European Molecular Biology Laboratory in Heidelberg (Germany),

CERTIFY:

That the research tasks who lead to the achievement of doctorate level titled: "Computer-Aided Protein Design for Engineering Enzymes to Recognize Specific Peptide and DNA Sequences", whose author is Mr.Emmanuel Fajardo Sánchez, have been done under our direction in the Institute of Molecular and Cellular Biology (IBMC) of the Miguel Hernández University of Elche (Spain) and in the European Molecular Biology Laboratory (EMBL) in Heidelberg (Germany).

In witness whereof and for all pertinent purposes, the present certificate is hereby issued in Elche, on 31st May 2007.

Signed:

Prof. Dr. Luis Serrano Pubull

Prof. Dr. Gregorio Fernández Ballester

UNIVERSITAS
Miguel
Hernández

“To know is so good!” Pedro Fajardo Martinez;

my Father, having just finished listening to the explanation about this scientific work.

“¡Que bueno es saber!” respuesta de mi padre tras acabar de escuchar la explicación sobre este trabajo científico.

A mi familia.

Acknowledgements

On the following lines I want to thank all the people who played a part in this PhD thesis.

Thanks to Prof. Dr. Antonio Vicente Ferrer Montiel, for believing in my enthusiasm for science and allowing me to enter the Molecular and Cellular Biology Institute of Elche (IBMC) and academic research.

Thanks to my thesis director in Germany, Prof. Dr. Luis Serrano Pubull, for receiving me in his research group at the European Molecular Biology Laboratory (EMBL), for enabling me to take part in the amazing scientific atmosphere at the EMBL, and for sharing with me part of his knowledge. I am also grateful to him for showing me that it is possible to make a living out of a "hobby" (even though most hobbies don't involve the implicit difficulties) and that science is a real daily job (several nights and weekends included), that is very rewarding when you discover something that - up until that moment - has not been seen by anyone else.

I would like to thank especially my thesis director in Spain, Professor Dr. Gregorio Fernández Ballester, who took me as his student guiding me with his wisdom, helping me to take the right way through the hard paths of computational biology. I also thank him for his advice in all other aspects of every day life during my first steps in Spain and later on in Germany, as well as my other mentor Professor Dr. José Antonio Encinar, who also took part in some adventures in Heidelberg.

I would also like to thank the "APOPIS" project, integrated in the Sixth Framework Program of the European Union, as well as Cellectis SA. for the financial support to this work and for their assistance in the patent application for the product resulting from it.

Next I would like to acknowledge all my colleagues who have taken part of this phase of my career, especially those who have shared more time with me and who showed their affection, both in Spain and in Germany. They know who they are and they also know that they are a part of my life. A special word to Dr. Yus (la bioquímica) and Dr. Isalan (my English friend, what what what…), whom I deeply thank for their patience and help during the "gestation" of this doctoral thesis, and who have demonstrated how they love and live science with passion. I extend this to all my friends who despite having no relation with this field of knowledge, have been present throughout my whole life

Many thanks to my family, who has always shown interest in all my "experiments", and at all times given their moral support. To my parents who have given me life and wished for me to succeed and reach as far as I could, whatever my field of

choice. To my sister “the Prof”, for showing me that good things come to you when you work hard. Last, but certainly not least, I am forever thankful to “la Tata” and “el Nino”, for being with me at all times supporting me, and without whom I would have never discovered my calling, nor would I have had the will to follow it.

Agradecimientos

Como es de bien nacido ser agradecido, en las siguientes líneas voy a dar las gracias a las personas y entidades que de algún modo u otro, han formado parte de esta tesis doctoral.

Gracias al profesor Dr. Antonio Vicente Ferrer Montiel, por creer en mi ilusión por la ciencia y permitirme entrar en el instituto de biología molecular y celular de Elx (IBMC) y en el mundo de la investigación académica.

Gracias a mi director de tesis en Alemania, el profesor Dr. Luis Serrano Pubull, por admitirme en su grupo de investigación, por dejarme disfrutar de la estupenda atmósfera científica que se siente en el laboratorio de biología molecular europeo (EMBL). Por transmitirme parte de sus conocimientos y por enseñarme que es posible vivir de lo que para nosotros los científicos, además de ser "una afición" más (pese a todas las dificultades implícitas), es un trabajo diario (con muchas noches y fines de semana incluidos) que se ve recompensado por el hecho de ser el primero, en descubrir algo nunca visto por nadie hasta ese momento.

A continuación agradezco especialmente a mi director de tesis en España, el profesor Dr. Gregorio Fernández Ballester, que me acogiera como su pupilo y me guiara con su sabiduría por el buen camino de los arduos senderos de la biología computacional. Agradezco además sus consejos en los demás aspectos de la vida cotidiana, durante los primeros pasos en España y en las andaduras por tierras teutonas junto a mi tutor el profesor Dr. José Antonio Encinar, que también participó de algunas aventuras puntuales en Heidelberg.

Agradezco además al proyecto "APOPIS", integrado en el Sexto Programa Marco de la Unión Europea y a Cellectis SA. la financiación de este trabajo y la ayuda prestada con la patente europea producida como resultado del mismo.

Seguidamente, me es gratificante recordar a todos los compañer@s que han sido parte de esta etapa de mi carrera, especialmente a los que han compartido más tiempo conmigo y me han demostrado su afecto tanto en España como en Alemania. Ell@s saben quienes son y que forman parte de mi vida. Mención honorífica para dos ejemplos de los anteriores, la Dra.Yus (la bioquímica) y el Dr.Isalan ("my English friend, what what what" Mark), a los que agradezco profundamente su paciencia y ayuda prestada durante "la gestación" de esta tesis doctoral, y que me han demostrado que aman y viven la ciencia con pasión. Ídem para el resto de amigos que aunque no tengan ninguna relación con este campo del saber, han estado presentes a lo largo de toda mi vida de estudiante y en la "extraescolar".

Muchas gracias, a mi familia que se ha interesado por "los experimentos" y me han dado apoyo continuamente. A mis padres, que me han dado la vida y siempre han querido que llegara lo más lejos posible en lo que eligiese hacer en ella. A mi hermana "la profe", por demostrarme que cada uno obtiene la recompensa merecida por su tesón y esfuerzo constante. Y en último lugar (aunque debería ser el primero), se lo agradezco infinitamente a "la Tata" y a "el Nino", porque han estado siempre y en todo momento apoyándome y sin ellos no habría descubierto mi vocación, ni habría tenido voluntad para seguirla.

UNIVERSITAS
Miguel
Hernández

Abstract

This dissertation focuses on the design of two enzymes, the Tobacco Etch Virus "Nuclear Inclusion a" (NIa) endoprotease (TEV protease) and the *Chlamydomonas reinhardtii* I-CreI homing endonuclease (I-CreI meganuclease). For this purpose FoldX was used, a protein design algorithm developed in our laboratory, which is based on physical and empirical parameters and which uses the protein structure to perform mutations and theoretical energy calculations.

The TEV protease recognizes and cuts specifically a canonical amino acid sequence, and is commonly used as a molecular tool in protein purification. The aim was to change the recognition site of this enzyme in order to direct the cleavage to specific sequences of interest, thus increasing its applicability.

Secondly, meganucleases are sequence specific dimeric endonucleases with large palindromic cleavage targets. The meganuclease I-CreI was designed to avoid the formation of homodimers and to favour the formation of obligate heterodimers. This approach enormously increases the repertoire of non-palindromic unique target sites on the genome that can be recognised by artificial enzymes. Such redesigned enzymes could be used in a wide range of applications, including the correction of mutations responsible for inherited monogenic diseases.

In summary, this thesis shows that computer-aided protein design is an effective tool in developing enzymes "a la carte" and has great potential for providing new molecular tools and biotherapies.

Resumen

Esta tesis se enfoca en el diseño asistido por ordenador de dos enzimas, la endonucleasa de Inclusión Nuclear a (NIa) del virus del grabado del tabaco (proteasa TEV) y una endonucleasa llamada meganucleasa I-CreI, descubierta en el alga verde unicelular *Chlamydomonas reinhardtii*. Para este propósito se ha usado FoldX, un algoritmo de diseño de proteínas desarrollado en nuestro laboratorio, que está basado en parámetros físicos y datos empíricos, y que utiliza la información estructural de la proteína para llevar a cabo las mutaciones y los cálculos teóricos de sus energías.

La proteasa TEV reconoce y corta específicamente una secuencia canónica de aminoácidos, y es comúnmente usada como herramienta molecular en diferentes técnicas entre las que destaca la purificación de proteínas recombinantes. Nuestro objetivo se enfocó en el cambio del sitio de reconocimiento de la proteasa TEV, para intentar dirigirla contra otras secuencias de interés específicas, y así incrementar su aplicabilidad.

Por otra parte, las meganucleasas forman dímeros que reconocen y cortan específicamente largas secuencias dianas en el ADN. Los monómeros de la enzima I-CreI, se rediseñaron para impedir la formación de homodímeros y permitir solamente la formación de heterodímeros obligados. De este modo, se incrementa enormemente el repertorio de dianas únicas no-palindrómicas reconocidas en el genoma. Estas nuevas enzimas pueden ser usadas en un amplio rango de aplicaciones, incluyendo la corrección de mutaciones responsables de enfermedades hereditarias monogénicas.

Por tanto, con este trabajo se demuestra que el diseño computacional de proteínas es una herramienta joven, aunque eficaz para rediseñar enzimas “a la carta”. Igualmente, el continuo desarrollo de la investigación y los conocimientos de estructura de proteínas, permitirá a este campo seguir evolucionando durante los próximos años.

TABLE OF CONTENTS

VI. MATERIALS & METHODS

ABBREVIATIONS

(aa/s)	Amino acid/s
(Aβ42)	Amyloid beta-protein
(BSA)	Bovine Serum Albumin
(CAD)	Caspase-Activated DNAse
(CGD)	Chronic Granulomatous Disease
(CPD)	Computational Protein Design
(DEE)	Dead-End Elimination
(DSBs)	DNA Double-Strand Breaks
(ΔG)	The free energy of unfolding
(EDTA)	Ethylenediamine tetra-acetic acid
(E:S)	Enzyme:Substrate
(FDA)	United States Food and Drug Administration
(GA)	Genetic Algorithm
(GFP)	Green Fluorescent Protein
(GST)	Gluthatione S-Transferase
(HEs)	Homing Endonucleases
(HEPES)	(4-(2-HydroxyEthyl)-1-PiperazineEthaneSulfonic acid)
(IPTG)	IsoPropyl-1-Thio-D-Galactopyranoside
(Kon)	Association rate
(MC)	Monte-Carlo simulation
(nAChRs)	Nicotinic acetylcholine receptors
(NCBI)	National Center for Biotechnology Information
(NC-IUBMB)	Nomenclature Committee of the International Union of Biochemistry and molecular biology
(NIa)	Nuclear Inclusion a endoprotease, TEV protease
(Occ)	Atomic occupancy
(ORF)	Open Reading Frame
(PASM)	Position Alanine Scoring Matrices
(PBS)	Phosphate Buffered Saline Buffer
(PDB)	Protein Data Bank. pdb; files with the structural information

(PEEF)	Physical Effective Energy Functions
(SCA)	Sickle Cell Anemia
(SCID)	Severe Combined Immunodeficiency Diseases
(SCMF)	Self-Consistent Mean Field
(SDS-PAGE)	Sodium DodecylSulfate-Polyacrylamide Gel Electrophoresis
(SEEF)	Statistical Effective Energy Functions
(Sfact)	Scaling factor
(SNPs)	Single Nucleotide Polymorphisms
(TEV)	The Tobacco Etch Virus
(wt)	Wild type

Amino Acid	**3-Letter**	**1-Letter**
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

UNIVERSITAS
Miguel
Hernández

I. INTRODUCTION

Proteins are the main players in cell physiology and behaviour. They not only catalyze almost all of the reactions that characterize carbon-based life, but also control virtually all biological processes. The L-isomers of the 20 amino acids that are found normally within proteins confer a vast array of chemical versatility. The precise amino acid content and sequence of each protein is determined by the sequence of the bases in the gene that encodes that protein. The chemical properties of the amino acids and particularly the tertiary/quaternary structure of proteins determine their biological activity.

In 1973 Anfinsen demonstrated that proteins contain the necessary information within their amino acid sequences to completely determine how they will fold into a native three dimensional structure. The stability of the resulting disposition corresponds to a global free-energy minimum (Anfinsen, 1973). The field of protein folding and stability has been a critically important area of research for years. Despite considerable progress in understanding the basic rules of secondary structure formation and protein stability, the well-known protein folding problem is far from being solved. Understanding how proteins establish their tertiary conformations, and interactions with other proteins, still remains as one of the great unsolved mysteries in biology today, even though it is currently being very actively investigated.

1.1 PROTEIN ENGINEERING

1.1.1 Overview

Through site-directed mutagenesis, it has been shown that by varying protein sequences, new structures and functions can be generated. The field started to be productive at the beginning of the 80´s, when it was noted that existing biological systems could be used to create nanoscale molecular machines with designed functions (Drexler, 1981). However, the big potential of this approach could only be fully realized

by the means of establishing reliable predictive methods, i.e. to predict which sequences will perform the desired functions. These strategies comprise the field known as protein design and the molecular methodology to carry them out is known as protein engineering.

1.1.2 General Strategies

There are three general strategies to perform protein design: rational design, directed evolution and computational protein design. All of them are connected, and can be used individually or synergistically in a global approach, depending on the available resources and the particular problem to solve.

Rational protein design consists of making desired changes in the amino acid sequence of the target protein, that are predicted to elicit the desired improvements of the new function (Baker and DeGrado, 1999; Bryson et al., 1995; Hellinga, 1997). This has the advantage of being relatively inexpensive and is based on modern well-developed drug discovery, target-based methods and site-directed mutagenesis techniques (Balakin et al., 2006). These approaches usually require structural knowledge about the target proteins to be redesigned and their small molecule ligands. Protein structural biology research is providing this information.

Actually, there are more than 43000 protein structures available in the protein data bank (PDB) http://www.pdb.org/ (Berman et al., 2007). Furthermore, when this structural knowledge of the target proteins and/or their small molecule ligands are unavailable, it is possible to use structural models inferred from a homologous protein structure by homology modelling (Ginalski, 2006).

The second strategy is known as directed evolution protein design. One of the most effective strategies in directed protein evolution is to accumulate mutations gradually, sequentially, or by recombination, using random mutagenesis, while applying

selective pressure. This is typically achieved by the generation of libraries of mutants followed by efficient screening of these libraries for a given function, by subsequent repetition of the process, using improved mutants from the previous screening. An additional technique known as DNA shuffling (Drexler, 1981; Harayama, 1998) mixes and matches pieces of successful variants in order to produce better results. This process involves the assembly of two or more DNA segments into a full-length gene by homologous, or site-specific, recombination. Before the assembly, the segments are often subjected to random mutagenesis by error-prone PCR or random nucleotide insertion.

The advantage of directed evolution techniques is that they require no prior structural knowledge of a protein, nor is it necessary to be able to predict what effect a given mutation will have. Indeed, the results of directed evolution experiments are often surprising in that desired changes are often caused by mutations that no one would have expected.

The main drawback is that they can require high-throughput techniques (Kurtzman et al., 2001; Harayama, 1998; Drexler, 1981), which are not generally available to academic labs. Because a large number of recombinant DNA molecules must be mutated, and the products screened for desired qualities, the total number of variants often requires expensive robotic equipment to automate the process. Furthermore, not all desired activities can be easily screened for. Because of all that, this approach is being mainly used by pharmaceutical companies to develop and improve therapeutic proteins.

This scientific work is based on the third strategy, namely computational protein design (CPD). Over the next chapter, CPD will be introduced, together with how this field is taking advantage of the huge advances in computational and scientific software during the last two decades. CPD is getting to be an important area of research, with constant progress being reported. The field has a very promising future.

1.2 Computational Protein Design

1.2.1 Overview

Taking into account the scenario mentioned above, computational protein design (CPD) is a mixture of rational and combinatorial protein design strategies, that use powerful computers and bioinformatic tools to generate new molecules. The *in silico* methodologies (Di et al., 2006) seek to identify low-energy amino acid mutated sequences for a specified target protein structure.

CPD was first conceived as the inverse of the protein folding problem, since its goal is to generate amino acid sequences that adopt a specific three-dimensional fold (Kurtzman et al., 2001; Harayama, 1998; Pabo, 1983). This approach usually utilizes the main-chain coordinates of a known protein structure as a fixed scaffold. Various amino acid types are modelled at each designed position (sequence space), and potential mutations are suggested, based on their interactions with the scaffold and with each other. Although the backbone is held fixed during a CPD calculation, various conformations of each amino acid type at each position are sampled to find sequences expected to stabilize the fold and satisfy any additional functional requirements.

1.2.2 Optimization Techniques

The distribution of energetically accessible conformations available to each amino acid side-chain is approximated using a set of discrete, low-energy conformations called rotamers (Janin and Wodak, 1978; Dunbrack, Jr. and Cohen, 1997; Shapovalov and Dunbrack, Jr., 2007). At the beginning of a typical CPD calculation, rotamers of the user-specified amino acid types are assigned to residue positions. The problem is thus to find a choice of rotamer at each position such that the fold is stabilized and the desired function is achieved.

For instance, the smallest polypeptide (L)-structure found in the Protein Data Bank (PDB) is the called α-Conotoxin Im1 with 12 residues (pdb:1IM1). This toxin is produced by predatory species of marine snails *Conus imperialis*. Conotoxins are potent antagonists of nicotinic acetylcholine receptors (nAChRs), ligand-gated ion channels involved in synaptic transmission (Rogers et al., 1999). Taking into account all the combinations of 20 amino acids permitted at this sequence space, even without attending to any rotamers, one would obtain $20^{12} = 4.096 \times 10^{15}$ sequence-structure solutions. But most proteins actually have hundreds of residues, rather than 12, and as mentioned above, to obtain accurate designs, it is also necessary to use a good rotamer library, which adds even more possibilities. Therefore, in real-life situations, the number of theoretically possible sequences to consider rapidly goes into astronomic figures. To perform this exploration in a reasonable time there are different optimization techniques ranging from deterministic procedures known as dead-end elimination (DEE) and Self-Consistent Mean Field (SCMF) optimization, to stochastic methods like Genetic algorithms (GA) or Monte-Carlo (MC) simulated annealing (Desjarlais and Clarke, 1998; Rogers et al., 1999; Voigt et al., 2000).

CPD has been considerably successful for modulating protein thermal stability (Blanes-Mira et al., 2001; Malakauskas and Mayo, 1998), designing protein cores (Desjarlais and Handel, 1995; Ventura and Serrano, 2004; Bolon et al., 2003; Ventura and Serrano, 2004), metal binding sites (Hellinga and Richards, 1991), enzyme-like biocatalysts (Bolon and Mayo, 2001), binding to DNA (Arnould et al., 2006; Ashworth et al., 2006), engineering complete proteins (Dahiyat and Mayo, 1997a; Dantas et al., 2003), changing protein-protein interaction affinity and specificity (Fernandez-Ballester and Serrano, 2006; Reina et al., 2002; van der Sloot et al., 2004; Kortemme et al., 2004; Baker and DeGrado, 1999), predicting binding targets of a particular fold at genome level (Kiel et al., 2005; Kolsch et al., 2007), studying folding mechanisms (Nauli et al., 2001), and new topologies (Harbury et al., 1998; Kuhlman et al., 2003), as well as the design of small-molecule protein receptors and indicators (Looger et al., 2003; Palmer et al., 2006). These achievements suggest that these techniques have already reached the point where

they can be applied and extended to modulate and engineer function in a biological context, by altering molecular recognition processes.

1.2.3 Accurate Energy Functions

Two main challenges are present in this field. First, as described above, the conformational problem and the sequence space have to be properly scanned. Second, the energy function used by the protein design software must be accurate enough to identify protein sequences with the desired three-dimensional conformation and the global free energy minimum.

Amino acid sequences and conformations are scored using a set of energy functions designed to reproduce the features of stable proteins. A wide range of strategies for estimating protein energies is used, from methods based on the statistical analysis of known protein structures on the one hand, to more physically based methods on the other. These were designated by Lazaridis and Karplus as Statistical and Physical Effective Energy Functions (SEEF and PEEF, respectively) (Lazaridis and Karplus, 2000). A third class of function is also widely used: Empirical Effective Energy Functions (EEEF) are based mainly on empirical data derived from experimental work on proteins (Mendes et al., 2002). Although the specific energy functions used, and their parameters, vary between different CPD implementations, most include a function that prevents atomic overlap and favors van der Waals interactions (Balakin et al., 2006; Dahiyat and Mayo, 1997b) and a function that benefits the formation of hydrogen bonds (Balakin et al., 2006; Dahiyat et al., 1997a; Morozov and Kortemme, 2005). Although interactions between a protein and its aqueous environment are crucial for stability, it would be prohibitively time-consuming to model water molecules explicitly in a CPD calculation (but not if you only predict water bridges as FoldX does; see 1.2.4.1 and Schymkowitz et al.). Therefore, solvation potentials are used to reward the burial of hydrophobic groups and to penalize the burial of polar groups; energies are computed using surface area

(Balakin et al., 2006; Lee and Richards, 1971; Marko et al., 2007; Iqbalsyah and Doig, 2005; Street and Mayo, 1998) or occluded volume models (Colonna-Cesari and Sander, 1990; Lazaridis and Karplus, 1999). Electrostatic interactions may be modeled using Coulomb's law with a constant dielectric (Lee et al., 2002; Zollars et al., 2006), a statistical pair potential (Kuhlman et al., 2003), or methods including multiple geometry-dependent dielectric constants (Wisz and Hellinga, 2003). These energy functions were designed to simulate different conformations of a single sequence and can give spurious results when used to choose between different sequences. Therefore, the scoring functions are typically supplemented with heuristic, statistical, and negative design terms to compensate for the limitations of the inverse folding model. These terms include heuristic estimates of side-chain entropy (Pokala and Handel, 2005), penalties for non-polar exposure (Dahiyat and Mayo, 1997b), statistical rotamer probabilities (Kuhlman et al., 2003), and composition-based unfolded state energies (Pokala and Handel, 2005; Kuhlman et al., 2003). Proteins have been successfully designed with multiple combinations of these functions, but no consensus has yet been reached on the ideal set of functions or the proper weight for each term (Gordon et al., 1999).

1.2.4 Force Fields

As mentioned above (see 1.2.2), different types of search algorithms are implemented with a molecular mechanics force field (Head-Gordon and Brown, 2003; Voigt et al., 2000). Various groups have developed modern protein force fields (Ponder and Case, 2003; Mackerell, Jr., 2004), depending on the design approach. For instance: AMBER (Sorin and Pande, 2005), ANLIZE (Stolworthy and Shirts, 1997), DEZYMER (Hellinga and Richards, 1991), DREIDING (Mayo et al., 1990), ECEPP/3 (Zhan et al., 2007), EGAD (Pokala and Handel, 2005), GROMOS (Christen et al., 2005), CHARMm (Mackerell, Jr. et al., 2000), METAL-SEARCH (Clarke and Yuan, 1995), OPLS (Jorgensen et al., 1996), ORBIT (Dahiyat et al., 1997b; Dahiyat and Mayo, 1997a) and others.

Force fields are usually composed of electrostatics, dihedral angle and van der Waals terms. Since ideal geometry is always assumed for protein design calculations, then bond-angle and bond stretching terms are not considered. Partial charges and dihedral angle parameters are derived from electron distributions from quantum theory. The parameters for van der Waals terms are determined from small-molecule crystal structures. These parameters are further adjusted by simulations that attempt to reproduce experimental data, such as small molecular crystal structures and heats of vaporization.

While these models work reasonably well for all atomic molecular dynamics simulations, they would require considerable modification for protein design calculations. Energies must be adjusted to reduce artifacts resulting from the use of discrete rotamers and fixed backbones. Energy terms that describe solvation must be added. A reference state needs to be defined, since the relevant value for protein design is the difference in energy between the folded and the unfolded states (Koehl and Levitt, 1999; Wernisch et al., 2000). Finally, these terms must all be weighted appropriately depending on the goal, and the overall computational time required must be taken into account.

1.2.4.1 FoldX

The FoldX (http://foldx.embl.de/) force field (FOLDEF) (Guerois et al., 2002; Schymkowitz et al., 2005b), developed in our research group, was used in this thesis. It was programmed to provide a fast and accurate estimation of mutational free energy changes on the stability of a protein, or a protein complex. It was also successfully applied to the prediction of protein folding pathways by removing most of the interactions between pairs of residues that are not in contact in the native state (the so called Gō-like models and their progeny approaches) (Guerois and Serrano, 2000). FoldX aims to describe the energetic contributions to protein stability in simple empirical terms that allow easy interpretation by non-specialists. It is thus geared at high-throughput structural biocomputing tasks, such as screening the effect of Single Nucleotide

Polymorphism (SNPs) on protein stability (Reumers et al., 2006) or *in silico* drug screening.

The FoldX energy function includes terms that have been found to be important for protein stability. The free energy of unfolding (ΔG) of a target protein is calculated using the equation:

$$\Delta G = W_{vdw} \cdot \Delta G_{vdw} + W_{solvH} \cdot \Delta G_{solvH} + W_{solvP} \cdot \Delta G_{solvP} + \Delta G_{wb} + \Delta G_{hbond} + \Delta G_{el} + \Delta G_{Kon} + W_{clash} \cdot \Delta G_{clash} + W_{mc} \cdot T \cdot \Delta S_{mc} + W_{sc} \cdot T \cdot \Delta S_{sc} + W_{scplx} \cdot T \cdot \Delta S_{scplx}$$

ΔG_{vdw} is the sum of the van der Waals' contributions of all atoms with respect to the same interactions with the solvent. ΔG_{solvH} and ΔG_{solvP} is the difference in solvation energy for apolar and polar groups respectively when going from the unfolded to the folded state. ΔG_{wb} is the extra stabilizing free energy provided by a water molecule making more than one hydrogen bond to the protein (water bridges) that cannot be taken into account with non-explicit solvent approximations (Petukhov et al., 1999). ΔG_{hbond} is the free energy difference between the formation of an intramolecular hydrogen bond compared to intermolecular hydrogen bond formation (with solvent). ΔG_{el} is calculated from a simple implementation of Coulomb's law, in which the dielectric constant is scaled with the burial of the bond under consideration. In order to improve the accuracy of the force field, hypothetical atoms are included in the calculations of the Coulombic interactions in order to capture some specific aspects of protein stability: i) Charged atoms are placed at the N- and C-terminal of each α-helix, to obtain some measure of the helix dipole interaction, and ii) Aromatic rings carry positive charges on the edges and negative charges above the centre of the ring.

ΔG_{Kon} is only applied to protein complexes, as the additional electrostatic contribution between atoms of different polypeptide chains and is based on the empirical equation of Schreiber et al., which was shown to give a good estimation of the association rate (k_{on}) of complex formation (Selzer et al., 2000). ΔG_{clash} provides a measure of the steric overlaps between atoms in the structure. ΔS_{mc} is the entropy cost for fixing the backbone in the folded state. This term is dependent on the intrinsic tendency of a

particular amino acid to adopt certain dihedral angles (Munoz and Serrano, 1994a). ΔS_{sc} is the entropic cost of fixing a side chain in a particular conformation (Abagyan and Totrov, 1994). Finally, ΔS_{scplx} is only applied on protein complexes and is the entropy cost of this complex formation, due to the loss of translational and rotational entropy upon complex formation.

The energy values of ΔG_{vdw}, ΔG_{solvH}, ΔG_{solvP}, ΔG_{hbond}, ΔG_{Kon} and ΔG_{clash} attributed to each atom type have been derived from a set of experimental data, but ΔS_{sc}, ΔS_{mc} and ΔS_{mcplx} have been taken from theoretical estimates. The terms W_{vdw}, W_{solvH}, W_{solvP}, W_{clash}, W_{mc}, W_{sc} and W_{scplx} correspond to the weighting factors applied to the raw energy terms. They all equal 1, except for the van der Waals' contribution which is 0.33 (the van der Waals' contributions are derived from vapor to water energy transfer, while in the protein they go from solvent to protein). W_{clash} varies depending on the use of FoldX: when analyzing point mutations this weight is relaxed, but when doing protein design this weight is fully applied.

**** Effect of Solvent Exposure***

Many experimental studies show that interactions at the surface of a protein usually contribute less to the stability of a protein than those in the core (Matthews, 1995; Serrano et al., 1992). This can be rationalized as an effect of increased flexibility at the protein surface in an environment close to that of the unfolded state. Therefore, an important part of the energy calculation is based on the inclusion of solvent effects in an implicit manner, except in the special case of water bridges. To estimate the solvent accessibility of a given atom, FoldX uses the solvent contact model (Gilis and Rooman, 1997), which considers the volume occupied by protein atoms around the target atom, called the atomic occupancy (*Occ*). The occupancy of a given atom *i* (Occ (*i*)) is the sum of the fragmental volumes of the atoms surrounding this atom within a threshold distance of 6Å (Janardhan and Vajda, 1998; Colonna-Cesari and Sander, 1990).

In FoldX, the atomic free energy of solvation, the van der Waals' interactions and the electrostatic interactions, together with the entropic terms, are scaled with respect to the atomic occupancies. As a first approximation, FoldX assumes that the strength of an interaction (solvation effects, van der Waals' or electrostatic) and the entropic cost for fixing the conformation of a residue should vary linearly with the atomic occupancy *Occ*(*i*).

For each atom *i*, the un-scaled energy terms are multiplied by the scaling factor (Sfact (i)) that is calculated from the atomic occupancy *Occ*(*i*) as :

$$Sfact\,(i) = \frac{Occ\,(i) - Occ_{\min}\,(t_i)}{Occ_{\max}\,(t_i) - Occ_{\min}\,(t_i)}$$

where $Occ_{\min}$ (*i*) and $Occ_{\max}$ (*i*) are the minimal and maximal occupancies of an atom of type t_i as estimated by Topham et al. (Topham et al., 1997).

**** Prediction of Water Binding Sites***

The modeling of water-protein interactions is critical for the accurate calculation of protein energy and their interactions with other macromolecules. A very accurate approach often used in molecular dynamics simulations is to consider a box of explicit water molecules during the calculations, which reproduces most aspects of solvation accurately (Baker, 2005). However, the technique is computationally extremely demanding (Mehta et al., 2004; Lee et al., 2004) and several approximations have been proposed to allow a higher computational efficiency (Feig et al., 2004; Baker, 2005). These approximations range from electrostatic models in which the waters are represented as simplified dipoles, often with restricted degrees of freedom to allow for

speedier calculations, to a range of continuum approximation methods based on the accessible surface area (Feig and Brooks, III, 2004). Because the solvent is treated as a uniformly distributed property surrounding the biomolecule of interest in these continuum methods, they cannot account for the formation of stable structures of water molecules near the surface or inside the protein (Petukhov et al., 1999). One such type of structure, that is often observed in high-resolution crystal structures, is called the water bridge, in which a water molecule forms a hydrogen bond with up to two donor and two acceptor groups on the protein. FoldX uses a continuum-solvation model based on experimental transfer energies of model compounds representing amino acids from water to vapor, and to cyclohexane, extended with the explicit consideration of water bridges (Guerois et al., 2002). This approach was pioneered by Petukhov et al. (Petukhov et al., 1999) and has been adopted by others (Jiang et al., 2005). It allows keeping fast calculations with more accurate energy evaluations. In some cases, the inclusion of structural waters improves the prediction of correct energy changes upon mutation (Jiang et al., 2005; Guerois et al., 2002).

**** Exceptions***

--*Polar Groups Desolvation*

For non-charged polar residues and non-charged polar backbone atoms the solvation penalty is calculated by first looking at the number of H-bonds they are making. If this number is equal to the maximum, then FoldX applies the full desolvation penalty independently of the solvent accessibility. If the number is lower, FoldX divides the desolvation penalty by the number of hydrogen bonds and subtracts this value, multiplied by the number of hydrogen bonds made from the desolvation value.

-- Chains Entropy Cost

For the main chain and side chain entropy which are calculated at the residue level and not at the atomic level, FoldX considers the mean value of the occupancies of the atoms that compose the main chain and the side chain respectively.

For residues not making electrostatic or hydrogen bond interactions from the side chain, FoldX corrects the entropy contribution by the Sfact (i). For the rest of the amino acids, the force field checks the electrostatic and hydrogen bond contributions and if these contributions are higher than the side chain entropic cost corrected by the Sfact (i), FoldX increases the entropic cost value to match the electrostatic and hydrogen bond contributions (provided that they are not higher than the maximum entropic cost value, otherwise it applies full entropy cost).

For the main chain, the algorithm applies the same correction as for the side chain entropy. The only difference is that FoldX looks also to the residues preceding and following the target residue. If those are making backbone hydrogen bonds then FoldX penalizes more the main chain entropy contribution of the target residue.

1.2.5 Designing New Molecular Tools and Therapies

In spite of all the challenges that CPD has to solve over the next years, already some great successes stand out. Recent work has focused on a small number of common biological functions, such as the interaction of proteins with nucleic acids, small molecules, and other proteins. There has been specific progress in engineering monoclonal antibodies (Presta, 2006), cytokines (Luo et al., 2002) and tumor necrosis factors (van der Sloot et al., 2006). Design approaches are also enabling the development of a new class of biotherapeutics, including hormones (Filikov et al., 2002) and viral fusion inhibitors, for the treatment of HIV and other viruses (Chan and Kim, 1998; Eckert

and Kim, 2001; Turner and Summers, 1999). In addition, major efforts toward the *de novo* design of catalytic activity are underway (Kaplan and DeGrado, 2004).

Several rational design and engineering strategies, such as docking or redesign of exposed hydrophobic residues, core residues, linear epitopes, loops, termini or binding sites, have been developed to improve properties such as solubility, stability, conformational control, immunogenicity, protease susceptibility, attachment of fusion partners or interaction affinity and specificity while controlling or improving desired biological activity (Marshall et al., 2003).

For example, stability is important throughout the production process and for the shelf-life of the final product, together with influencing the pharmacokinetic and dynamic properties of the protein therapeutic (Lazar et al., 2003). Several strategies, including both rational design (van den et al., 1998; Pantoliano et al., 1989; Villegas et al., 1996; van der Sloot et al., 2004) and directed evolution methods (Giver et al., 1998; Jung et al., 1999) are nowadays successfully used to improve stability of proteins (Fersht and Winter, 1992; van den and Eijsink, 2002). An inconvenience in a rational approach is that only a limited number of potentially improved designed variants can be tested. By contrast, directed evolution methods, like phage display or two-hybrid screening, allow large numbers of variants to be generated and tested. However, suitable selection or screening procedures are required, which are often not available, or need very intensive efforts.

It is worth mentioning outstanding examples, such as the redesign of calmodulin into a variant with eight mutations that maintained high affinity for the target peptide, while showing decreased binding affinity for non-targeted peptides (Shifman and Mayo, 2002; Shifman and Mayo, 2003; Shifman et al., 2004). Specificity design was taken a step further when a PDZ domain mutant was engineered to bind a new sequence (Reina et al., 2002).

More recent work on protein-protein, protein-peptide and protein-DNA interaction specificity has revealed the value of negative design. Negative design is the process by

which undesired properties are considered and designed against. A conceptually straightforward implementation of negative design is one in which the scoring function favors mutations that are stabilizing in the target structure while destabilizing in alternative structures. As in the redesign of coil-coil dimers (Havranek and Harbury, 2003), successful designs selected amino acids that were predicted to favor the target dimer state over the rest. Further evidence of the usefulness of negative design was demonstrated in the redesign of the colicin E7 DNase-Im7 immunity protein interface (Joachimiak et al., 2006).

The use of the FoldX force field (see 1.2.4.1) on protein design has been already successfully applied on a variety of proteins (van der Sloot et al., 2004; Reina et al., 2002; Kiel et al., 2004; Kiel et al., 2005; Kempkens et al., 2006; van der Sloot et al., 2006; Musi et al., 2006; Kolsch et al., 2007; Kolsch et al., 2007; Kolsch et al., 2007; Villanueva et al., 2003; Fernandez-Ballester et al., 2004), including the new redesign of a protease cleavage site, and the redesign of a meganuclease dimer interface, that are presented here. These kinds of results allow us to foresee a central role for CPD methods in biotechnology and in protein therapy optimization.

Hereafter, the two main parts of this dissertation will be introduced. First of all, the *in silico* exercise to redesign a protease that specifically cleaves proteins which contain a known canonical sequence will be described. Second, we will discuss the redesign of the protein-protein interface of meganuclease homodimers to make an obligatory heterodimer that cleaves very specifically a non-palindromic DNA sequence.

1.3 Proteins Cleaving Proteins

1.3.1 Overview of Proteases

Proteases are enzymes that are indispensable for all forms of life. They account for 2% of the genomes of most organisms (including humans), and they control the activation, synthesis and turnover of proteins (Puente et al., 2003). Proteases are pivotal regulators of many physiological processes during conception, birth, growth, maturation, aging and death (Blobel, 2000; Maymon et al., 2000; Whitcomb and Lowe, 2007).

Proteases are also essential for the replication and transmission of viruses, parasites and bacteria that cause infectious diseases. On the other hand, they are very important in homeostasis, apoptosis and host defense. Over the last decade, proteases have generated considerable biomedical interest owing to the identification of several human pathologies in which these enzymes are implicated, including neurodegenerative disorders, inflammatory diseases and cancer (Balbin et al., 2003; Coussens et al., 2000; Gutierrez-Fernandez et al., 2007; Mohammed et al., 2004; Camins et al., 2006). For instance, the proteasome (an intracellular multicatalytic protease complex) present in eukaryotic cells is mainly responsible for selective degradation of intracellular proteins involved in the execution of key cellular functions (Mukhopadhyay and Riezman, 2007). Thus, proteasome inhibition is actually a potential therapeutic target in cancer and inflammatory diseases. Actually, both the EMEA (European Medicine Evaluation Agency) and the FDA (United States Food and Drug Administration) granted approval for the use of some proteasome inhibitors for the treatment of relapsed multiple myeloma. At present, several phase II and phase III trials are ongoing in solid tumors and hematological malignancies. This inhibition could result in the stabilization and accumulation of proteasome substrates, a phenomenon that may act in inducing signals in cells such as cell cycle arrest and activation of apoptotic programs (Zavrski et al., 2007).

Thus, over 560 human proteases are forming our degradome (Lopez-Otin and Overall, 2002; Overall et al., 2004), and it is now known that single amino acid mutations in at least 10% of human regulatory proteases result in hereditary/genetic diseases (Puente et al., 2003; Puente et al., 2005). Therefore, a few proteases have already been studied and targeted by pharmaceutical companies and the academic community, who have successfully developed selective and non-toxic drugs for the treatment of HIV/AIDS (Hui et al., 1991; Martin et al., 2003), stroke and coronary infarction (Berg et al., 2003) and other diseases. An ever-growing number of protease inhibitors are now entering clinical trials.

1.3.2 Classification of Proteases

Attending to proteolytic action, they can be divided into two different categories:

1) Limited proteolysis; in which a protease cleaves only one or a limited number of peptide bonds of a target protein, leading to the activation or maturation of the formerly inactive protein. For example, conversion of pro-hormones to hormones (MacGregor et al., 1976).

2) Unlimited proteolysis; in which proteins are degraded into their amino acid constituents. The proteins to be degraded are usually first conjugated to multiple molecules of the polypeptide ubiquitin. This modification marks them for rapid hydrolysis by the proteasome in an ATP-dependent process. Another pathway consists in the delivery to lysosomes. Proteins transferred into these protease-rich compartments undergo rapid degradation (Terman et al., 2006).

The Nomenclature Committee of the NC-IUBMB (International Union of Biochemistry and Molecular Biology) classify enzymes by the reactions they catalyze (http://www.chem.qmul.ac.uk/iubmb/enzyme/). However, they have used the chemical nature of the enzyme in certain cases, where classification based on specificity is difficult. Also, this commission has recommended using the term *peptidase*

for the subset of peptide bond hydrolases (*subclass E.C 3.4.*). The more widely-used term protease is synonymous with peptidase, proteinase and proteolytic enzyme, and will therefore be used in this dissertation.

Proteases comprise two groups of enzymes: the endopeptidases, which cleave peptide bonds at points within the protein and the exopeptidases, which remove amino acids sequentially from either the N or C-termini.

The endopeptidases are divided into five sub-subclasses: serine, cysteine, aspartic, metallo- and threonine endopeptidases, depending on the basis of catalytic mechanism.

The MEROPS database, (http://merops.sanger.ac.uk/), is a very complete information resource for proteases and the proteins that inhibit them (Rawlings et al., 2006). It uses a hierarchical, structure-based classification of the peptidases. Here, each peptidase is assigned to a Family on the basis of statistically significant similarities in amino acid sequence, and families that are thought to be homologous are grouped together in a Clan.

1.3.3 TEV Protease

1.3.3.1 Biological Context

The tobacco etch virus (TEV) belongs to the potyviridae family (Brunt, 1992), which is a subfamily of a large group of positive-strand RNA viruses that are responsible for a number of plant and animal diseases (Ryan and Flint, 1997). A single open reading frame of the TEV RNA genome (NCBI, NC_001555) encodes the TEVgp1 polyprotein of about 346 kDa, that is subsequently proteolytically processed into more than a dozen individual proteins (Dougherty et al., 1989a), three of which are proteases: P1, HC-Pro and NIa peptidases (Carrington and Dougherty, 1987b; Carrington and Dougherty, 1987a). At the initial stage, all of them are autocatalytically released from the polyprotein N-terminus, but only NIa participates in all the subsequent stages of proteolysis (Parks et al., 1995).

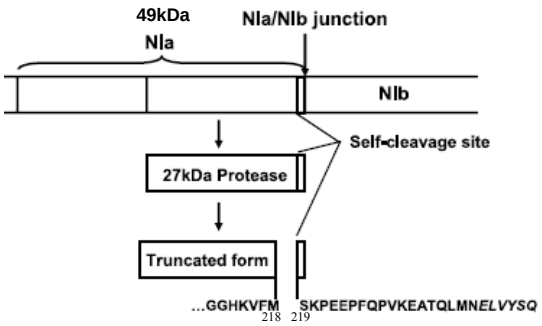


Figure 1.1 Diagram of TEV polyprotein processing, resulting in release of the 27 kDa Nuclear Inclusion a protease (NIa). *The diagram also shows the site of the self-cleavage. The last six residues in the sequence of the 27 kDa protease are shown in italics to emphasize that these six residues, present in the current structure, correspond to the internal proteolytic site. (Adapted from (Nunn et al., 2005).*

1.3.3.2 Features and Cleavage Site

Nuclear Inclusion a protease (NIa) has a molecular mass of 49 kDa and consists of two subunits. Its C-terminal fragment of 27 kDa, is sufficient for proteolytic activity (see Fig.1.1). During the late stages of infection, it has been shown to exist as an independent protein, widely known as TEV protease (Carrington and Dougherty, 1987a; Carrington and Dougherty, 1987b). This catalytic domain is a cysteine endopeptidase; three residues are implicated in the catalysis: Asp81, His46 and Cys151 (see Fig.1.3).

It was established that TEV protease is homologous to picornavirus 3C proteases (Gosert et al., 1997; Lawson and Semler, 1991) and structurally similar to serine proteases, such as trypsin and chymotrypsin; however, it utilizes the cysteine151 thiol group, instead of a serine hydroxyl as the active nucleophile (Dougherty et al., 1989b; Gorbalenya et al., 1989). Thus, TEV protease is inhibited by thiol alkylating reagents such as iodoacetamide (Rzychon et al., 2004) and some detergents (Mohanty et al., 2003).

TEV protease recognizes and cleaves with high specificity a seven amino acid canonical sequence (Dougherty et al., 1989a), Glu-X-X-Tyr-X-Gln//(Ser/Gly), where X can be various amino acyl residues, although the heptapeptide consensus is ENLYFQ//G (Kapust et al., 2002; Kapust et al., 2001) (see Fig.1.2). Cleavage occurs between the conserved Gln and Ser residues (slashes), over a broad temperature range (Carrington and Dougherty, 1988).

Although the specificity of TEV protease towards these amino acid sequences is very high, it is not absolute as demonstrated by Dougherty and co-workers, who made an *in vitro* characterization of each position of the TEV protease cleavage site (Dougherty et al., 1988; Dougherty et al., 1989a). It was also shown that the enzyme can cleave itself at the bond Met218–Ser219 producing a truncated protein with a significantly lower activity (Kapust et al., 2001). This autolysis phenomenon was observed only *in vitro* and has been a subject of several studies, and a number of stable mutants of this enzyme have been proposed (Parks et al., 1995; Kapust et al., 2001).

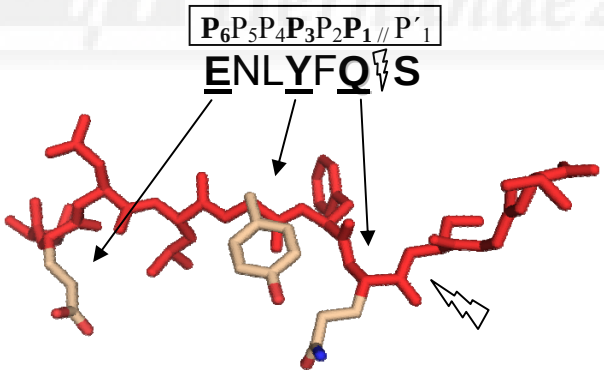


Figure 1.2 TEV protease substrate. *The peptide substrate with the most common canonical sequence cleaved by TEV protease. The key positions P_6, P_3 and P_1 are underlined. The P_2, P_4, P_5 and $P`_1$ are more permissive positions (see Fig.1.3). The lightning indicates the cleavage site.*

1.3.3.3 Structure

Two point mutant structures are available for TEV protease. One of the mutated structures is TEV-S219D, which is an active enzyme stable to autolysis. Another mutant, TEV-C151A, is inactive, since the catalytically active Cys151 residue is replaced by Ala. Both proteins have been crystallized as complexes (Phan et al., 2002; Nunn et al., 2005), with either the product of the proteolytic reaction, in the case of TEV-S219D (PBD code: 1LVM; 2.70Å resolution), or the intact substrate in TEV-C151A (PDB code:1LVB; 2.20Å resolution). There is also a full-length inactive mutant structure, in the absence of peptide, with the C terminus of the protease bound to the active site (Nunn et al., 2005) (PDB code: 1Q31; 2.70Å resolution).

The TEV protease structure consists of two domains in the form of antiparallel β-barrels, with important residues that compose the catalytic triad (His46, Asp81, and Cys151) located at the interdomain interface (see Fig.1.3). The enzyme does not appear to have been perturbed by the mutations in either structure, and the modes of binding of the substrate and product are virtually identical in these mutants.

Coming back to the canonical substrate sequence, the importance of each one of the six positions could be explained by taking into account the TEV protease binding pocket. This is formed by five sub-pockets, carrying most interactions corresponding to the key positions P_6, P_3 and P_1. TEV protease does not have sub-pocket for P_5 which explains why practically any residue can occupy this position with almost no impact in the cleavage efficiency (Dougherty et al., 1989a). Furthermore, experimental data (Kapust et al., 2002) has shown the P`1 pocket allows residues with short aliphatic side chains (Ser, Gly, Ala, Met and Cys), although this is partially exposed to solvent rather than completely buried within the complex. As mentioned, the main sub-pockets are building intricate hydrogen bonds and hydrophobic interaction networks that explain why this substrate is so specific for TEV protease. Altogether, this structural information provides the necessary guidance for re-engineering the enzyme to improve or alter its target site.

Figure 1.3 Active TEV protease structure. *Backbone representation of the 1.8 Å resolution crystal structure of TEV protease bound to the product. The characteristic two-domain antiparallel β-barrels folds are depicted in ribbon yellow and the amino acids of the active catalytic triad are shown in sticks. The product is shown in dark blue.*

1.3.3.4 Commons Uses and New Perspectives

Given the properties of TEV protease, this enzyme is used as a common tool for various site-specific proteolysis approaches. The simplest and perhaps the most well-known application of TEV protease, is the removal of affinity tags in purifications of recombinant proteins (Cabrita et al., 2006; Kim et al., 2005; Wawersik et al., 2005; Knuesel et al., 2003; Haspel et al., 2001; Eisenmesser et al., 2000; Kapust and Waugh, 2000; Melcher, 2000; Urabe et al., 1999; Smith and Kohorn, 1991; Parks et al., 1994; Waugh, 2005; Rigaut et al., 1999; Puig et al., 2001). Furthermore, this protease can also be used for various *in vivo* applications such as the inactivation of essential proteins, the mapping of functions to specific domains, as well as genetic screening procedures (Shih et al., 2005; Eser et al., 2007; Wehr et al., 2006).

To conclude the introduction about TEV protease, it is worth noting that this enzyme is an attractive target for protein redesign, to cleave new target sites. To begin with, it would be interesting to broaden or modify the field of action, compared to the specificity for the canonical sequence. Also to improve the wild-type specificity could be extremely useful. These features would open new possibilities for using this enzyme as molecular tool.

Another challenge could be to customize the protease against a well known target sequence of one disease-causing protein (Ross and Poirier, 2004). Examples of such disease proteins include the amyloid beta-protein (Aβ42), which is implicated in Alzheimer's disease (Christensen, 2007), and the expanded polyglutamine proteins, implicated in the pathogenesis of nine inherited neurodegenerative diseases, including Huntington's disease, various spinocerebellar ataxia types, dentatorubral pallidoluysian atrophy, and spinobulbar muscular atrophy (Mitsui et al., 2006). A successful redesigning would pave the way for this enzyme to be used as a therapeutic protein.

1.4 Proteins Cleaving DNA

1.4.1 Overview of Nucleases

Numerous types of DNases and RNases have been isolated and characterized. They differ, amongst other things, in substrate specificity (DNA or RNA respectively), cofactor requirements, and whether they cleave nucleic acids internally (endonucleases), or from the ends (exonucleases), or whether they attack in both of these modes.

Nucleases play an important role in the pathogenesis of various diseases. For instance the angiogenins, members of the pancreatic RNases superfamily, are implicated in cancer (Strydom, 1998). Also some DNases (as for example CAD; Caspase-Activated DNAse) are involved in programmed cell death (apoptosis) (Counis and Torriglia, 2006).

1.4.2 Definition of Meganucleases

By definition, meganucleases are sequence-specific endonucleases with large cleavage sites (12-45 bp) that can deliver DNA double-strand breaks (DSBs) at specific loci in living cells (Thierry and Dujon, 1992). They can be used to achieve very high levels of gene targeting efficiencies in mammalian cells and plants (Choulika et al., 1995; Donoho et al., 1998; Paques and Duchateau, 2007; Puchta et al., 1996; Rouet et al., 1994). Indeed, meganuclease-induced recombination is an efficient and robust method for genome engineering. The major limitation so far was the requirement for the prior introduction of a meganuclease target site in the locus of interest.

1.4.3 Homing Endonucleases

In nature, meganucleases are essentially represented by Homing Endonucleases (HEs), a widespread family of endonucleases including hundreds of proteins (Chevalier

and Stoddard, 2001). HEs are implicated in a process known as "Homing" (see Fig.1.4); these HEs are encoded by genes with mobile self-splicing introns (Dujon et al., 1989). After transcription, the internal open reading frame (ORF) results in expression of the endonuclease. Subsequently, the meganuclease binds a very specific target sequence (homing site) on the DNA of the host gene by making a double strand break. The similarity of sequence between both genes, allows a homologous recombination event that duplicates the mobile DNA into the recipient locus.

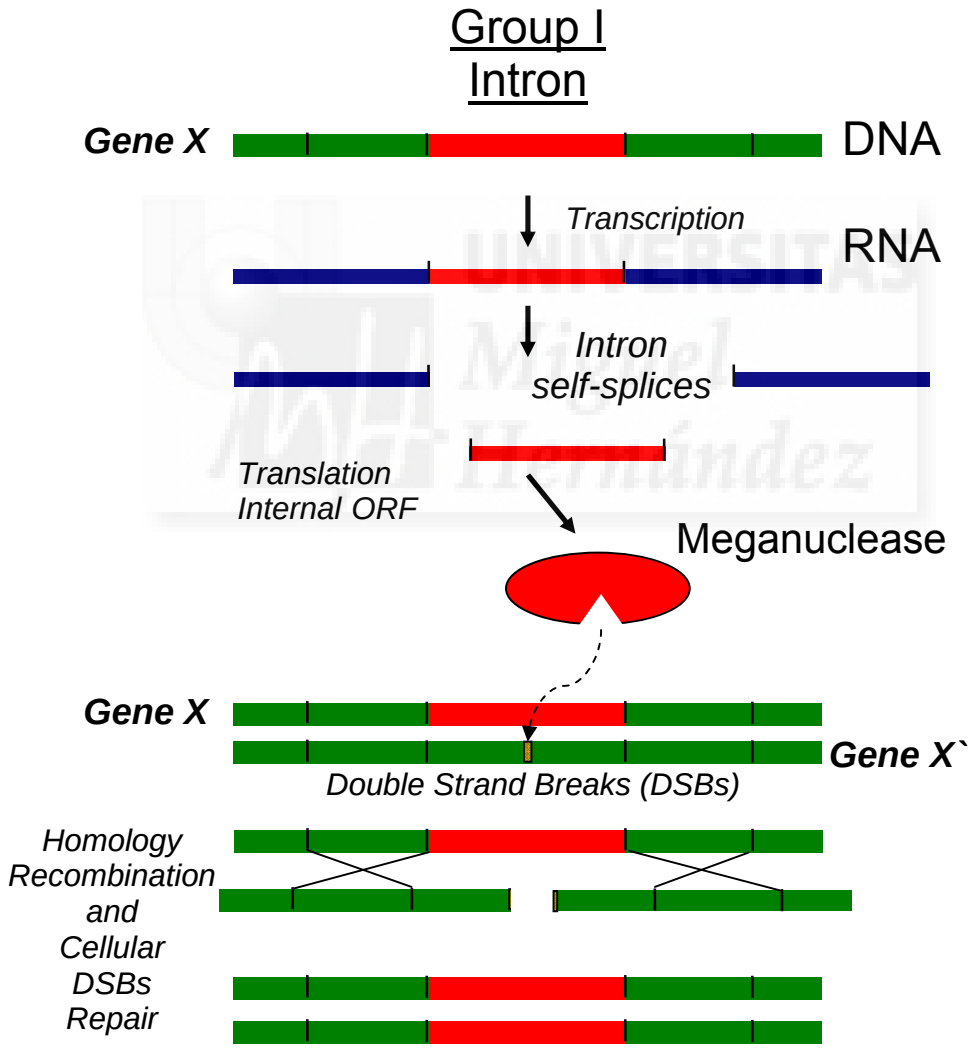


Figure 1.4 Homing mechanism of group I introns. *The intervening sequence of gene X is duplicated in its target allele, gene X'. See details above.*

Given their natural function, and their exceptional cleavage properties in terms of efficacy and specificity, HEs could provide ideal scaffolds to derive novel endonucleases for genome engineering. Data have accumulated over the last decade, allowing a relatively good characterization of the LAGLIDADG family, the largest of the four HEs families (Chevalier and Stoddard, 2001).

1.4.3.1 LAGLIDADG Family

LAGLIDADG refers to the only amino acid sequence actually conserved throughout the family, and is found in one or (more often) two copies in the protein. Proteins with a single motif, such as I-CreI (group I intron-encoded HE) (Heath et al., 1997), form homodimers and cleave palindromic or pseudo-palindromic DNA sequences, whereas the larger, double motif proteins, such as PI-SceI are monomers and cleave non palindromic targets.

Nine different LAGLIDADG proteins have been crystallized, showing a very striking core structure conservation that contrasts with the lack of similarity at the primary structure level (Bolduc et al., 2003; Chevalier et al., 2003; Chevalier et al., 2001b; Ichiyanagi et al., 2000; Jurica et al., 1998; Moure et al., 2002; Moure et al., 2003; Nakayama et al., 2006; Silva et al., 1999; Spiegel et al., 2006). In this core structure (see Fig.1.5), two characteristic αββαββα folds, contributed by two monomers, or two domains in dimeric LAGLIDAG proteins, are facing each other with a two-fold symmetry. DNA binding depends on the four β strands from each domain, folded into an antiparallel β-sheet, and forming a saddle on the DNA helix major groove. The catalytic site is central, formed with contributions from helices of both monomers. In addition to this core structure, other domains can be found; for instance the intein PI-SceI has a protein splicing domain, and an additional DNA-binding domain.

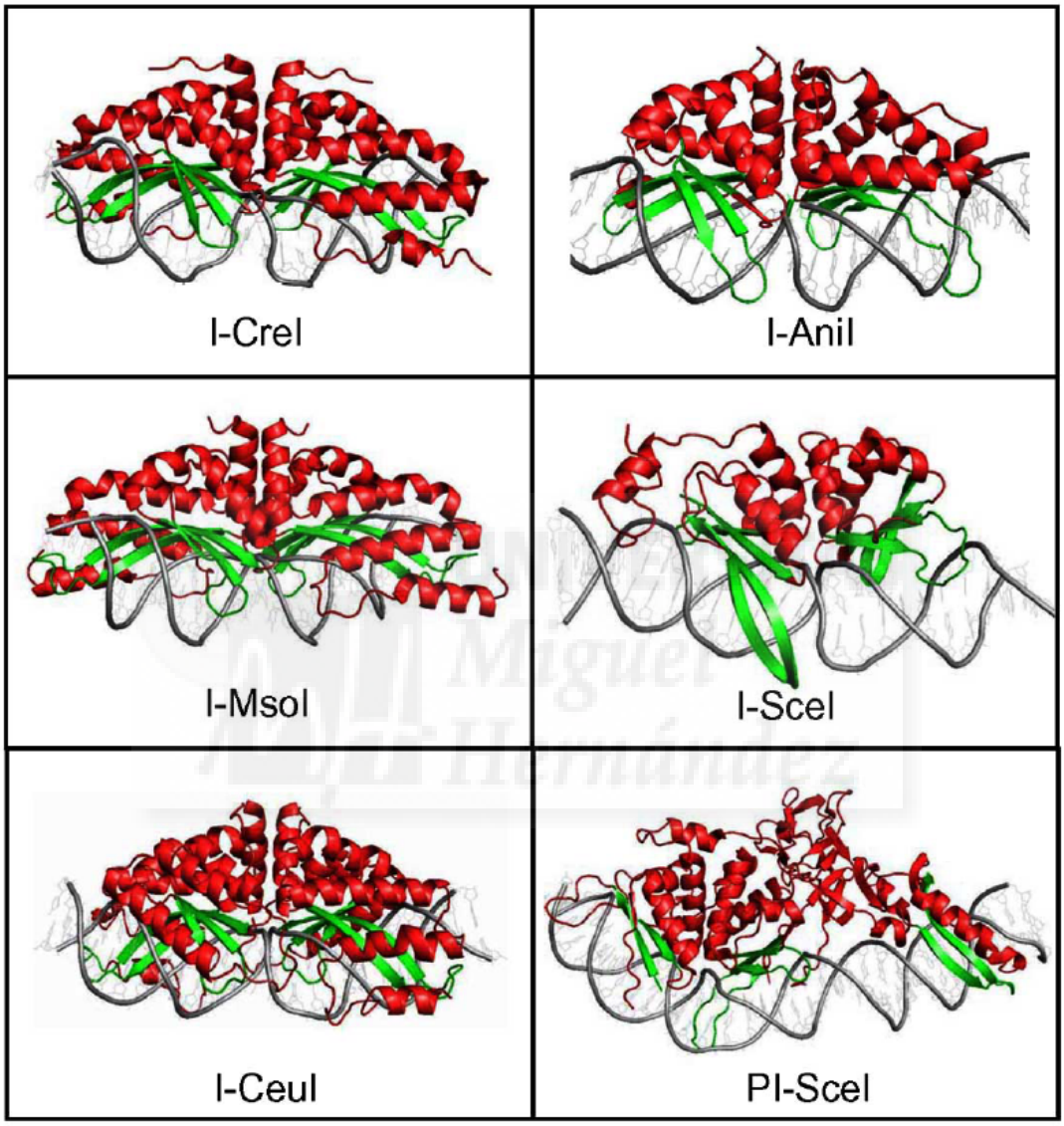


Figure 1.5 Meganucleases from the LAGLIDADG family. *The panel displays three homodimers on the left (I-CreI, I-MsoI and I-CeuI), and three monomeric proteins on the right (I-AniI, I-SceI and PI-SceI) bound to their DNA target. The αββαββα fold characteristic of this family is represented as a cartoon. In this very compact structure, the catalytic domain in the center is embedded in the two DNA binding interfaces formed by the αββαββα folds. Monomeric proteins have retained the symmetric structures of homodimers, but the sequences from their two moieties are extremely divergent. PI-SceI, an intein, has an additional protein splicing domain, which includes a supplementary binding interface.(Adapted from (Paques and Duchateau, 2007).*

1.4.3.2 Engineering Meganucleases

The extensive structural conservation within the meganuclease family has encouraged the mutagenesis and construction of chimeric and single chain HEs, which can withstand extensive modifications (Chevalier et al., 2002; Epinat et al., 2003; Steuer et al., 2004). Seligman and co-workers used a rational approach to substitute specific individual residues of the I-CreI αββαββα fold, and they could observe substantial cleavage of novel targets (Seligman et al., 2002; Sussman et al., 2004). The same kind of approach was applied to I-SceI recently by another group (Doyon et al., 2006). In a similar way, Gimble and coworkers modified the additional DNA binding domain of PI-SceI, and could obtain variant proteins with altered binding specificity (Gimble et al., 2003). Recent works made by our collaborators (see Fig.1.6) have shown that it is possible to obtain a large number of locally altered variants of the I-CreI meganuclease that recognize a wide new range of targets (Arnould et al., 2006), and to use and assemble them by a combinatorial process, to obtain entirely redesigned mutants with chosen specificity (Smith et al., 2006). These variants can be used to cleave genuine chromosomal sequences and open a wide range of applications, including the correction of mutations responsible for inherited monogenic diseases (Paques and Duchateau, 2007).

**** Meganuclease Design Challenges***

A limiting factor that still remains for the widespread use of the I-CreI meganuclease is the fact that the protein is a homodimer. Thus, although there is experimental evidence that mixing two meganucleases, that target two different DNA sequences, can result in the formation of a heterodimer that recognizes a hybrid DNA sequence (Arnould et al., 2006; Smith et al., 2006), this still results in a mixture of three different enzymes, including both homodimers (Arnould et al., 2006). Thus, redesigning the enzyme to obtain a pure obligatory heterodimer could solve the problem. This approach was carried out during this scientific work and will be presented in the next chapters.

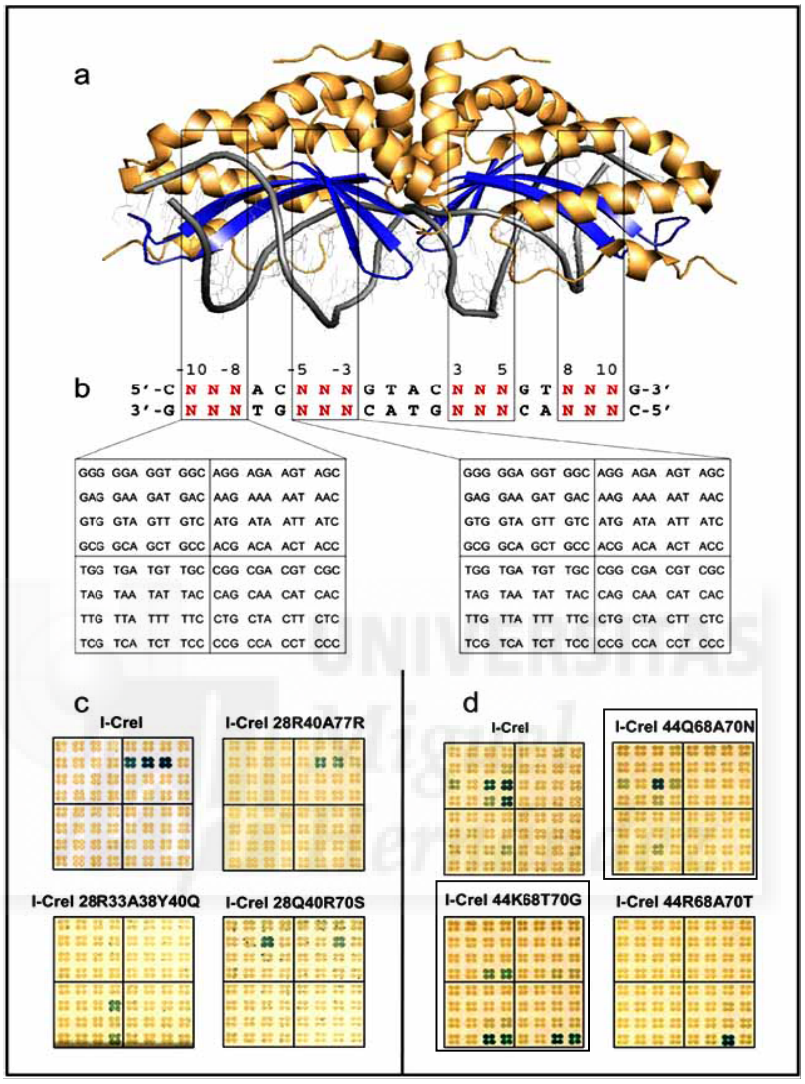


Figure 1.6 Local engineering of the I-CreI endonuclease. *(a) Distinct regions of the protein (boxed) can be engineered separately. Homodimeric mutants are generated, in order to cleave novel palindromic targets. As a consequence, two out the four boxed regions are engineered simultaneously. (b) Locally altered targets. Locally engineered region correspond to two pairs of symmetric DNA triplets, in position ±3, ±4, ±5 and ±8, ±9, ±10. Randomization of each pair of triplets results in a series of 64 different palindromic targets. (c and d) I-CreI variants with novel specificities. Residues in the vicinity of each triplet are randomized, and the resulting homodimeric mutants are screened against the 64 related targets, using a functional screening assay in yeast cells. Cleavage of a target by the mutant results in the restoration of a functional β-galactosidase gene by homologous recombination, and in blue staining in the presence of X-Gal. The profile of a few mutants is represented. Each mutant is tested against a series of 64 targets, differing from the I-CreI palindromic target by position ±8, ±9, ±10 (c) or ±3, ±4, ±5 (d). Squares mark the best I-CreI mutants (QAN and KTG). Targets are ordered as in (b). (Adapted from (Paques and Duchateau, 2007).*

1.5 Protein Design: Tools and Therapies

One of the most important applications of protein engineering is molecular therapeutics. Two approaches can be carried out: the removal of the damaged protein or the repair of the mutated genetic source.

First, in the case of well known misfolded proteins, such as Aβ42 (as discussed previously), that a very specific customized TEV protease could "clean" the aggregates from the extracellular space.

Second, meganuclease-induced recombination could be used for the correction of mutations responsible for monogenic inherited diseases such as Severe Combined Immunodeficiency Diseases (SCID), Sickle Cell Anemia (SCA) or Chronic Granulomatous Disease (CGD) by using gene therapy (Paques and Duchateau, 2007).

1.5.1 Gene Therapy

Gene therapy consists of the introduction of new genes into cells for the purpose of treating disease by restoring or adding gene expression. For instance, numerous growth factors and other proteins with the ability to promote the regeneration of tissues in the locomotive system have been identified, but in general their clinical use is often hindered by delivery problems. In principle, these problems can be overcome by delivering the relevant genes, as the therapeutic substances thereby can be persistently produced directly by local cells at the site of diseases.

1.6 Summary

Both structural biology and structural genomics initiatives are determining an increasing amount of structural information for proteins and protein complexes, each time with better resolution, and all are on-line and freely available. This makes more and more proteins directly amenable to redesign and provides templates with high sequence identity for structurally similar targets. This in turn allows the creation of accurate homology models that can be used as scaffolds for design.

Moreover, recent advances in *de novo* protein design and protein-protein docking will eventually allow the prediction of more accurate structural models of target complexes. These developments will allow successful designs with a reduced amount of false positive and false negative predictions, making the CPD field an important resource for developing new molecular tools and therapies.

UNIVERSITAS
Miguel
Hernández

II. OBJECTIVES

The main objective of this scientific work is to redesign new enzymes "a la carte", and more specifically to modify protein-protein interactions in order to change affinity and specificity in protein complexes and enzyme-peptide interactions.

To accomplish this goal, we will use structural information available in the Brookhaven Protein Data Bank and FoldX, a computational protein design algorithm developed by our group. The general methodology used can serve as a starting platform to tackle any other redesign of interest.

The particular objectives are:

1- The redesign of TEV protease to change its canonical recognition site.

The binding site of TEV protease will be redesigned to change the specificity for the key residues of the canonical target-site. An *in silico* screening strategy will be developed to search for optimal sequences and to determine the theoretical energy of protein-protein interactions, to discriminate the best solutions for each key position in the canonical recognition site of TEV protease.

2- To redesign the interaction interface of the I-CreI meganuclease homodimers to make obligate heterodimers.

Herein, by using the three-dimensional information for I-CreI meganuclease, the interaction surface between the monomers of this enzyme will be redesigned to facilitate heterodimerisation and at the same time to prevent the formation of homodimers, or at least make them thermodynamically unstable and thus to obtain an obligatory heterodimer.

The long term objective is to use these modified enzymes as molecular tools to open the way for new therapies. For example, TEV proteases could be redesigned to destroy disease-causing peptides such amyloidogenic proteins (e.g. Tau or Aβ42 proteins). Likewise, redesigned meganucleases could be used to repair monogenic disease genes such as those involved in SCID.

UNIVERSITAS
Miguel
Hernández

III. RESULTS

3.1 Redesigning TEV protease Specificity

3.1.1 Computational Screening and Redesign of the Recognition Sites

3.1.1.1 Global Design to Cleave any Substrate

To redesign the binding site of TEV protease, the X-ray structure of the inactive TEV protease (C151A) bound to its canonical substrate target sequence was used (PDB:1lvb, 2.20 Å resolution). The aim was to change the substrate specificity of the enzyme to bind and cleave a desired target sequence. As mentioned above (see Fig.3.1), there are three key residue-positions in the canonical sequence: P_6 (Glu302), P_3 (Tyr305) and P_1 (Gln307). The cleavage happens between P_1 and $P`_1$ (Ser or Gly) and the rest of the positions are more permissive and can contain other residues without affecting the binding of the substrate.

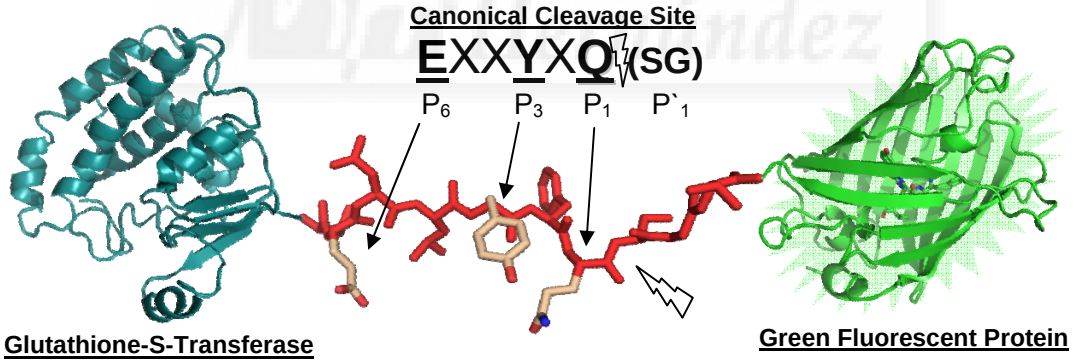


Figure 3.1 Representation of the substrate-reporter. *Between the Glutathione-S-Transferase (GST) and the Green Fluorescent Protein (GFP) tags, is the widely-used TEV protease Canonical Cleavage Site (ENLYFQ/S) with the key positions underlined. The X indicates that P_2, P_4 and P_5 are more permissive positions, tolerating many different amino acids. The lightning indicates where the cleavage occurs.*

First, the specificity range of the wt TEV protease was assayed, for binding to different substrates, without making any modification to the enzyme. FoldX was used to make an *in silico* positional scanning of the substrate to construct the pattern of substrate

tolerance. Considering the results published by Dougherty and co-workers, using *in vitro* assays to see the effect of mutations on each position of the substrate (Dougherty et al., 1989a), and comparing this with the *in silico* screening results obtained with FoldX, the general pattern looks as follows:

	P6	P_5	P_4	***P3***	P_2	***P1***	$P`_1$
in vitro pattern:	[EQPS]-	[DNHTC]-	[LIHTY]-	[YVK]-	[FCI]-	[QCGNF]-	/-[SINR]
in silico pattern:	[EHKR]-	[FVLSE]-	[RIHV]-	[YHK]-	[FSH]-	[QFTEN]-	/-[SALI]

"A la carte" overall pattern:

P6	P_5	P_4	***P3***	P_2	***P1***	$P`_1$
[EQHKRSP]-	[DNFVHLTSCE]-	[LIHRTVY]-	[YVHK]-	[FCSIH]-	[QCERTNF]-	/-[SANILR]

Therefore the rational design exercise was centred on altering the pockets on the TEV protease to accommodate new residues in these key positions. To achieve this, a mixed strategy of *in silico* testing and *in vitro* verification was carried out.

The global strategy was as follows:

First of all, the structure 1lvb.pdb was repaired using FoldX vs 2.65 (see 6.1.1.3). The resulting structure was used as template (wt TEV protease), for all the following steps on the global redesign (see Fig.3.2).

The second step was to perform an *in silico* screening of consensus positions (P_1, P_3 and P_6), by means of FoldX. The calculations were made using only 18 out of 20 natural amino acids (Gly and Pro were not included), while neighbouring positions in the protease were mutated to Ala.

Third, the resulting models were calculated for energy, to filter out those residues making strong clashes as well as non-sense solutions (see pattern with good solutions below). Other non-consensus positions in the substrate (P_2, P_4 and P_5) were first checked visually to estimate which residues were capable of fitting with minor modifications in the protease. These positions were also scanned *in silico* and the resulting models were

evaluated in terms of energy. In general, the exposed positions accepted almost all the natural amino acids that were tried, whereas other less exposed positions accepted only hydrophobic (or suitable polar) residues.

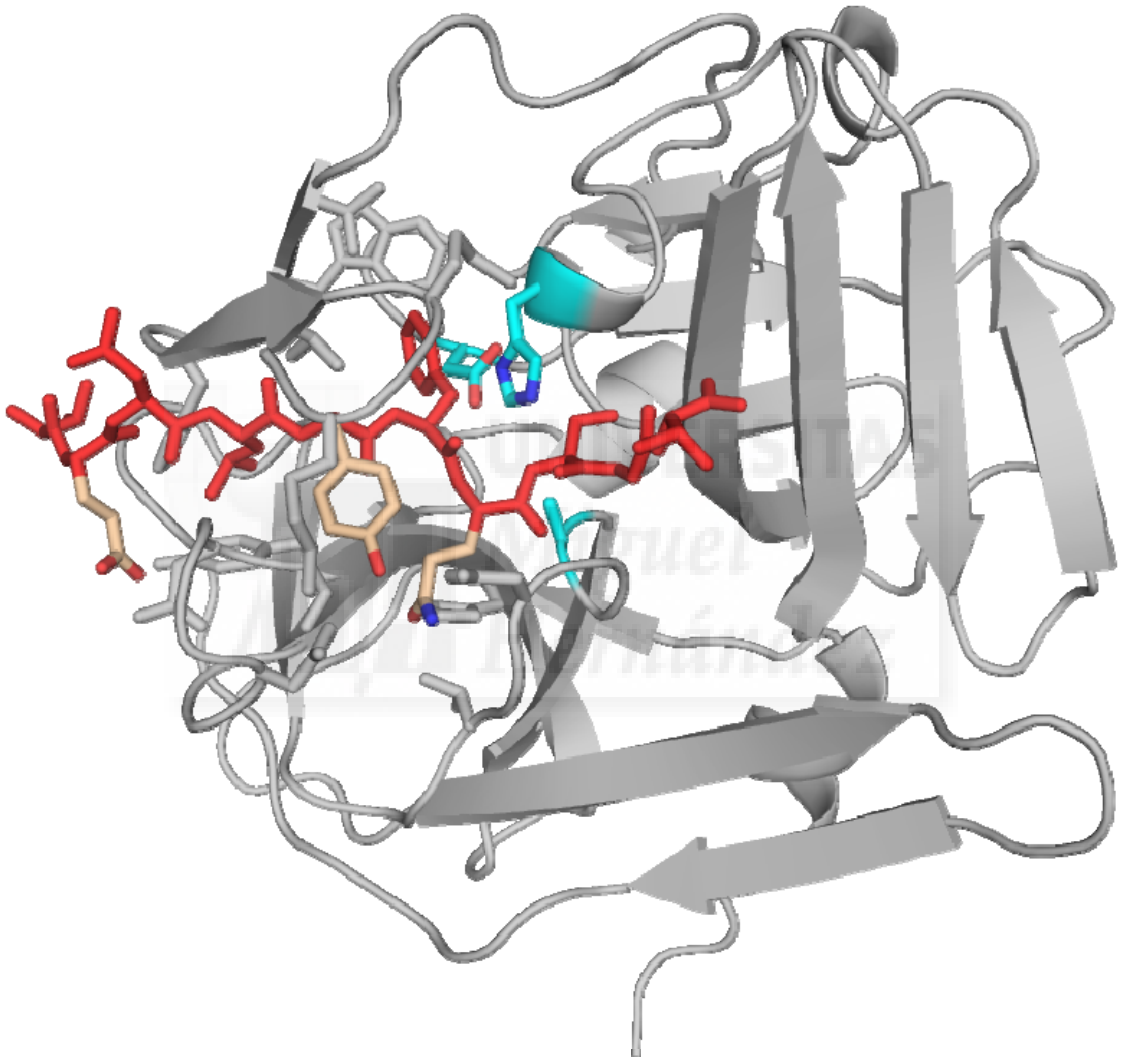


Figure 3.2 TEV protease structure. *Cartoon representation of the repaired structure of TEV protease (in silver) binding to the substrate (in red). The TEV protease residues forming part of the binding pocket are represented as sticks. The three key residues on the substrate are in salmon colour. The residues of the inactive (C151A) catalytic triad are in light blue.*

There was an immediate problem for the global design strategy described above. To get some usable information from this approach, it would have had to deal with thousands of combinations, that would require searching for single solutions for all positions in the protease, which in turn would require thousands of millions of possible combinations. In order to tackle this objective with the available resources, it was mandatory to dissect the problem into a series of very much smaller jobs, taking into account only one key position at a time.

3.1.1.2 Design of Position 307 (P_1)

The structural characterization of the binding pocket made by Phan and co-workers, shows the main interactions between the canonical substrate and the wt TEV protease (Phan et al., 2002). A plot with these interactions (see Fig.3.3) was made using the software LIGPLOT (Wallace et al., 1995). P1 is probably the most important position in the substrate since its correct binding determines the correct positioning of the P`1 to locate the peptide bond to be broken, in the position near the active site. In addition, the residue P5 (another key position) is also involved in the correct location of the P1 residue. Thus, it was decided first to focus the redesign of the TEV protease to bind modified sequences at position Q307.

The computational screening started with the mutagenesis of position P_1 to the 18 natural amino acids, while the other positions in the substrate remained unaltered. The neighbouring positions in the protease were mutated to Ala to avoid strong clashes. These positions were then tested to find the best combinations. For this, an exhaustive search in the sequence space was done to ensure that the best solutions were found, to hold the new environment in the P_1 pocket.

Since there are 4 (or 5) neighbouring positions interacting with the protease, the number of combinations to scan was still huge (20^4 or 20^5). For this reason it was necessary to split the problem into smaller jobs, so that the computational time could be

affordable. Therefore, for each single model of position P_1, a positional scanning of the individual positions in the protease was performed to construct position alanine scoring matrices (PASM), while the other selected positions remained as Ala. These PASM are based on the binding energy measurements of the resulting complexes and allow the rapid identification of amino acids or groups of amino acids that fit better in a given position.

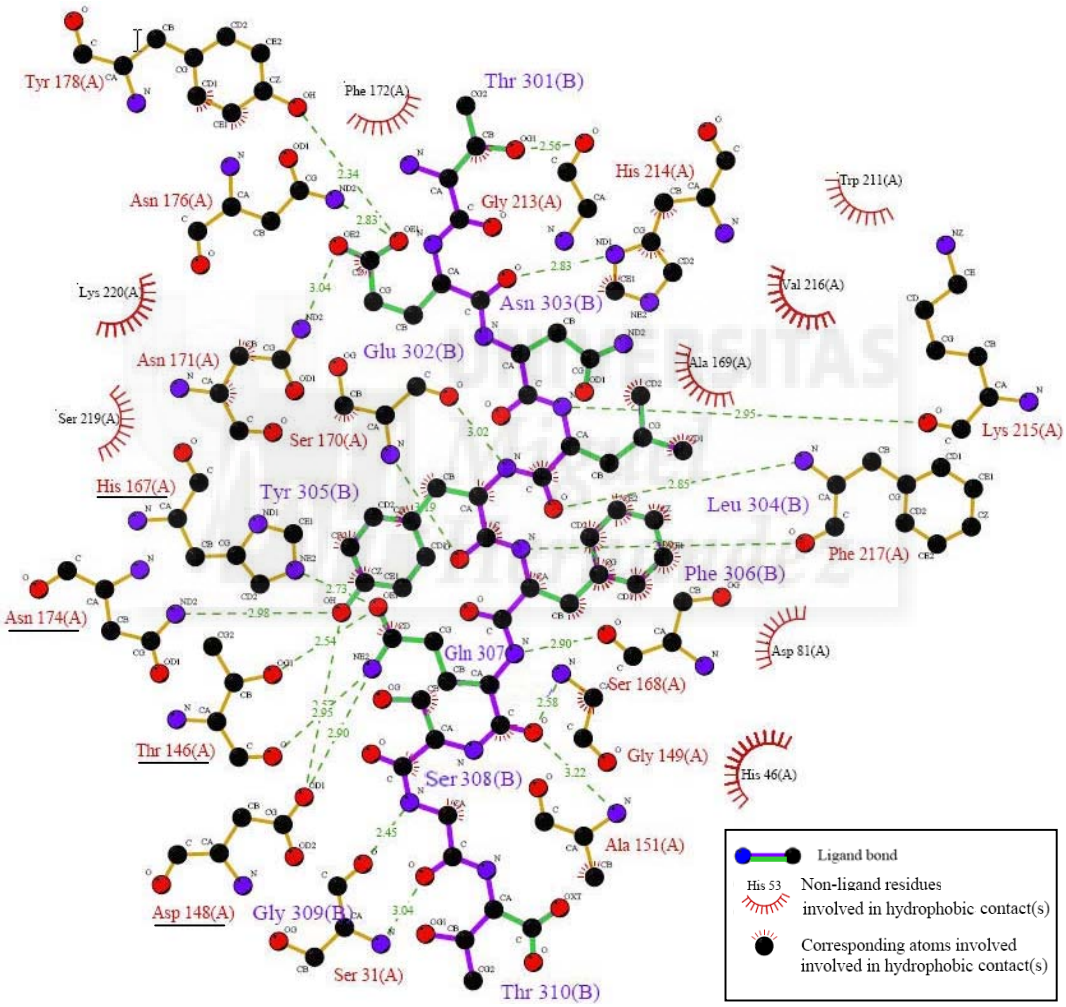


Figure 3.3 Plot representation of the TEV binding site. *Ligplot was used to depict the interactions between the canonical substrate TENLYFQGT (Residues 301-310, chain B) and the inactive TEV protease binding pocket (chain A). The residues implicated in the P_1 pocket are underlined.*

The position alanine scoring matrices (PASM), together with an additional visual inspection of the models, help to discriminate the most favourable amino acids per position (clustered by physical properties: polar, charged, hydrophobic, aromatic, and also large, small, etc.). This procedure also removes non-sense solutions, thus reducing the number of residues per position and consequently reducing the number of combinations.

All the positional scans were calculated by FoldX running over a cluster from the European Molecular Biology Laboratory (see 6.1). The evaluation of the binding energies of all the models obtained, allowed the construction of normalized PASM, one for each individual position involved in binding. Their analysis allowed the selection of up to five residues with good energies as candidates for the following steps of design (see Table 3.1).

SUBSTRATE P_1 POSITION MUTATION	REDESIGNED TEV PROTEASE RESIDUES					COMBINATIONS
	T146	D148	H167	S170	N174	
Q307A	ARNDC	DEH	H	-	RNDCE	75
Q307C	CILMN	DEHKN	RVHYK	HRWYF	AKN	1875
Q307D	MSTV	ST	RMLFI	-	MKLHQ	200
Q307E	T	NQKRH	H	-	NQKRH	25
Q307F	RDCQI	ADCEH	HFS	-	ARNCH	375
Q307H	NSAVL	QSTHK	YFLIM	NTVIP	-	3125
Q307I	CNQTV	DEQHN	LVHI	RHN	NK	600
Q307K	ITAMV	DER	FHLY	-	NH	120
Q307L	NCQIL	ADQEL	ANDCQ	-	NAR	375
Q307M	TACNS	EDK	HFLM	S	N	60
Q307N	ACVNT	EDNH	HY	-	NH	120
Q307Q	*TRNCQ*	*DR*	*HR*	*SRHN*	*NDCE*	*320*
Q307R	ATV	EDNQ	HFY	-	NH	72
Q307S	CNSV	DEN	FHL	FRHN	NK	288
Q307T	CNSV	DE	HLV	RHN	NK	144
Q307V	TQSNV	DE	VHL	-	NRK	90
Q307W	CQIKM	NDCQE	QHLM	-	ANDCQ	500
Q307Y	RNCQI	ANDCQ	QHST	-	RNCQE	500

Table 3.1 The top residues for each position of TEV protease interacting with each amino acid on the P_1 position of the substrate. *The five best candidates are highlighted in pale green and the best one is marked in deep green. The right column shows the number of combinations needed to cover all possibilities designs to be calculated.*

3.1.1.3 The Choice of the New Target

At this stage single models were obtained, containing one of the L-amino acids at position P_1 interacting with four (or five) neighbouring positions in the TEV protease, their corresponding PASM, and a selection of the best residues (up to five) per position. For instance for P_1 position carrying an Asp, the candidate designs have been narrowed down to 200 (see Table 3.1). Therefore, the number of combinations was now affordable and designs could be made on all positions in the neighbourhood of the P_1 pocket at a time.

In order to choose a mutation at position Q307 to be used to redesign the TEV protease and to be verified experimentally, the following factors were taken into consideration:

- The chemical properties of the residue; glutamine is a polar amino acid with no net charge and mildly hydrophobic properties. Therefore the new residue could be a charged.

- The physical properties of the residue; the glutamine side chain has two carbons and one amide group. Therefore the new side chain could be shorter.

- Experimental data in the literature pointed out which residues show very poor or no detectable cleavage when inserted in this position.

Ideally, the residue had to be similar in size but with different chemical properties. Aspartic acid is smaller than glutamine and is negatively charged and therefore chemically different to the wild-type substrate. In addition, this residue made the substrate not cleavable in the *in vitro* assays of wt TEV (Dougherty et al., 1989a). Therefore this choice was expected to allow the mutant TEV protease designs to discriminate well between wild-type and mutant substrate. Thus Q307D was chosen as

the substrate target, first for the *in silico* designs and subsequently for the experimental work to test the resulting mutant TEV proteases.

3.1.1.4 Designing TEV protease for the Q307D substrate

The following steps were undertaken to perform the design of the protease against the Q307D substrate.

First, the TEV protease model with the aspartate in the 307 substrate position (Q307D) was generated, and the interacting positions of the enzyme were mutated to alanine (T146A, D148A, H167A and N174A), using FoldX. This template was used to mutate each of the four individual positions to the 20 amino acids, leaving the remaining positions with alanine and giving rise to 80 new models (see Fig.3.4).

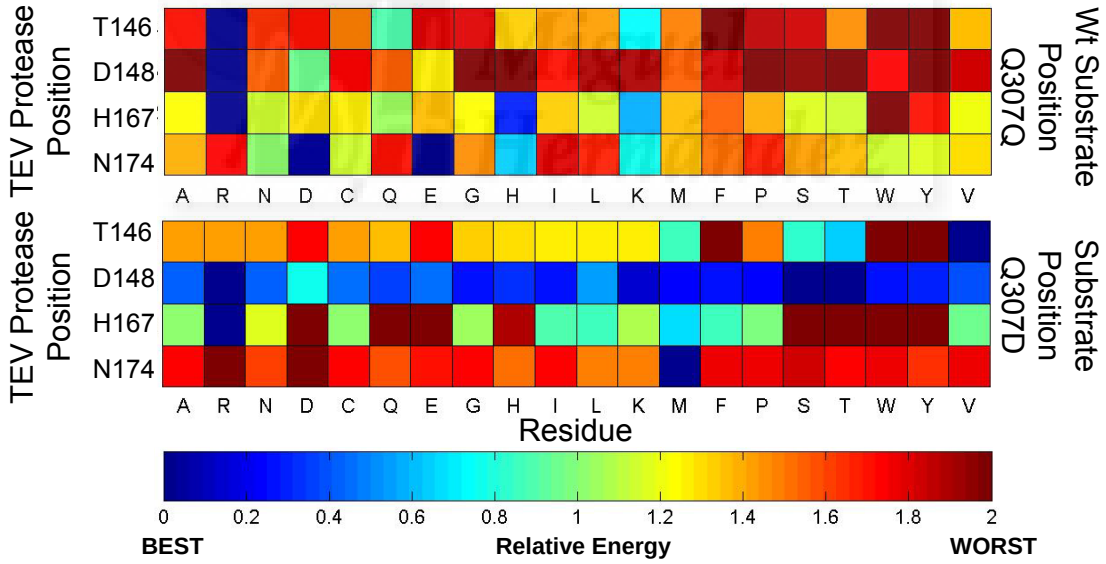


Figure 3.4 Position Alanine Score Matrix (PASM) of TEV protease's sub-pocket P_1 binding the wt substrate and the Q307D substrate. *The color status bar indicates the favorability of the energy (kcal/mol), relative to the best energy sequence, for each residue in the four positions while the other three positions are held as alanine.*

Second, these models were evaluated in terms of energy, and both the Gibbs free energy (ΔΔG) of binding interaction and the sum of intra-molecular clashes and ΔΔG of binding were normalized to construct the PASM. The PASM was used to discriminate which amino acids would be used in combination with the other positions (see table 3.1 and Fig.3.4).

Third, FoldX performed the mutagenesis combining all allowed residues in their respective positions, producing 200 new models and the corresponding energy values.

The table 3.2 shows the top 30 results of this analysis. They are ranked according to the energies given by FoldX. The classification was based on the binding ΔΔG and intra molecular clash ΔΔG. The stabilization energies were also taken into account to exclude very unlikely structures.

					Stability Energy (kcal/mol)		*Intraclash Energy (kcal/mol)*		*Binding Energy (kcal/mol)*	
	wt TEV template and wt substrate				-64.887		8.225		-36.022	
	Four alanines TEV Template and Q**307D** substrate				-59.307		8.224		-32.210	
RANK	Q307D substrate TEV Protease T146	D148	H167	N174	*Total Stability*	ΔΔG	*Total Intraclash*	ΔΔG	*Total Binding*	ΔΔG
1	V	T	R	K	-67.845	-8.537	10.206	1.982	-44.136	-11.926
2	T	T	R	K	-73.232	-13.924	9.126	0.902	-43.862	-11.652
3	T	S	R	K	-68.176	-8.869	8.949	0.725	-43.208	-10.997
4	V	S	R	K	-69.224	-9.916	9.608	1.384	-42.863	-10.653
5	S	S	R	K	-67.348	-8.040	9.179	0.955	-42.612	-10.402
6	M	T	M	K	-68.225	-8.917	8.478	0.254	-41.684	-9.473
7	M	T	R	K	-57.539	1.769	9.680	1.457	-41.627	-9.417
8	S	T	R	K	-66.620	-7.313	8.995	0.772	-41.625	-9.415
9	M	T	L	K	-64.831	-5.524	8.217	-0.007	-41.307	-9.096
10	M	T	F	K	-3.345	55.963	8.241	0.017	-40.899	-8.689
11	M	S	L	K	-62.822	-3.515	8.318	0.094	-40.684	-8.474
12	S	T	F	K	-68.234	-8.926	8.376	0.152	-40.653	-8.443
13	M	S	R	K	-60.569	-1.262	8.613	0.389	-40.639	-8.428
14	T	T	L	K	-67.867	-8.559	9.672	1.448	-40.461	-8.251
15	S	S	F	K	-69.126	-9.818	8.358	0.134	-40.412	-8.202
16	V	T	R	H	-67.768	-8.461	10.009	1.785	-40.297	-8.087
17	V	T	R	Q	-67.129	-7.821	10.059	1.835	-40.263	-8.053
18	T	S	I	K	-65.325	-6.017	10.988	2.764	-40.234	-8.023
19	M	S	M	K	-66.478	-7.170	8.354	0.131	-40.189	-7.979
20	M	S	F	K	0.437	59.745	8.303	0.079	-40.172	-7.962
21	V	T	M	K	-69.844	-10.536	8.323	0.099	-40.055	-7.845
22	T	S	L	K	-66.354	-7.046	10.371	2.148	-40.054	-7.844
23	V	S	L	K	-69.846	-10.539	8.716	0.492	-40.035	-7.824
24	V	S	M	K	-70.027	-10.719	8.543	0.319	-40.028	-7.818
25	S	T	I	K	-64.489	-5.182	10.097	1.873	-39.916	-7.706
26	S	T	L	K	-67.722	-8.415	8.346	0.122	-39.903	-7.692
27	V	S	F	K	-69.814	-10.506	8.970	0.746	-39.864	-7.654
28	M	T	I	K	67.084	126.39	11.516	3.292	-39.817	-7.606
29	S	S	L	K	-68.190	-8.883	8.483	0.259	-39.767	-7.557
30	S	S	I	K	-64.143	-4.835	10.402	2.179	-39.687	-7.476

Table 3.2 Top 30 designs. *The best energy examples of the 200 TEV protease designs for Q307D substrate. The energies of the repaired TEV protease are in dark green and the energies of TEV protease template with the four positions with alanines are in ochre.*

Moreover, a visual inspection of the best structures showed the globally preferred amino acids per position:

Position H167: This position is presented first because it is the major binding determinant and will hence govern the other positions. Arg fits quite well since its side chain is buried in the deep hole of the P1 pocket, and its guanidinium group establishes hydrogen bonds with Asp307 and to Asn177, thus compensating the polar groups (see Fig.3.5). In addition, Lys174 also accepts a hydrogen bond from Asp307. As an alternative, position T167 could hold a Leu, thus partly filling the hydrophobic hole in combination with Met146. In this case, a salt bridge with Asp307 is established mainly through Lys174 and one hydrogen bond with Ser148 (see Fig.3.5 Panel C).

Position N174: Here a Lys was introduced in order to provide a salt bridge to Asp307, as mentioned above.

Position D148: For this position Ser or Thr were selected to help to stabilize Tyr305 (position P_3 of the substrate) as well as Asp307.

Position T146: Met and Ser showed good energy values. However, Met seemed to fit better, filling the hydrophobic hole when a hydrophobic residue (Leu) was present in position 167. On the other hand, when Arg was modeled in position 167, Ala was chosen rationally in position 146, to avoid any molecular clash or steric hindrance, even though the energy values for Ala in this position were not among the best ones.

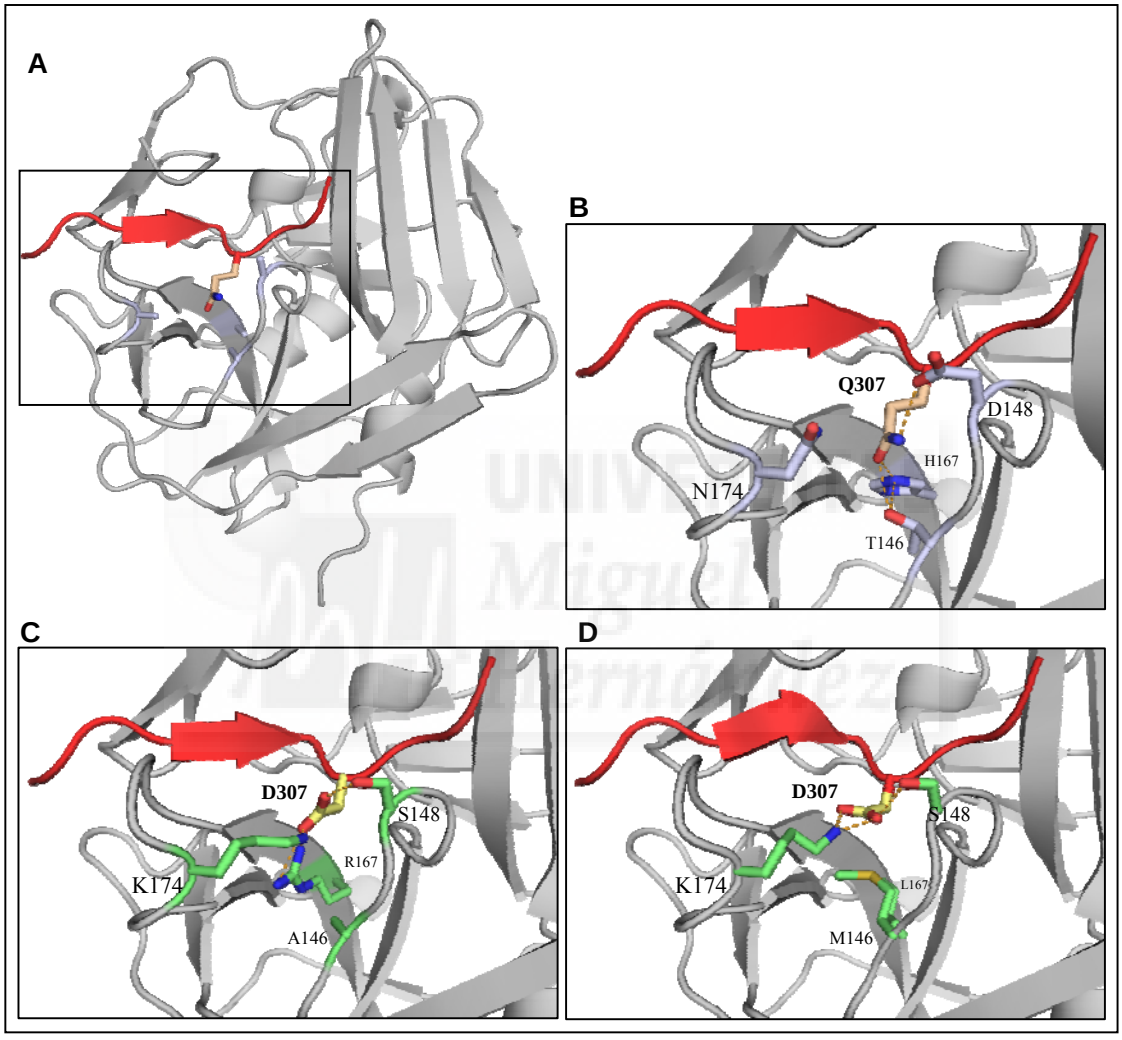


Figure 3.5 Detail of the designed P1 pocket. *(A) Cartoon representation of the structure of TEV protease binding to the substrate (in red) with the glutamine307 and the four residues implicated in the design mutated to alanines (in sticks); the box frames the binding pocket. (B) Zoom-in view of the P_1 binding pocket showing the original residues. (C) The best design: ASRK TEV protease. (D) The design MSLK TEV protease.*

Finally, after analyzing these energies and doing a visual re-evaluation of the best candidates, it was decided to produce, purify and test the following four proteases *in vitro*. It was expected that these proteases would recognize and cleave the new substrate Q307D (see Table 3.3 below).

	Q307D substrate TEV Protease				*Stability Energy (kcal/mol)*	*Intraclash Energy (kcal/mol)*	*Binding Energy (kcal/mol)*
Wt =>	T146	D148	H167	N174	-59.307	8.224	-32.210
Designs					ΔΔG	ΔΔG	ΔΔG
1	A	S	R	K	-5.246	0.075	-8.326
2	A	T	R	K	-7.765	0.194	-8.356
3	M	S	L	K	-3.515	0.094	-8.474
4	M	T	L	K	-5.524	-0.007	-9.096

Table 3.3 The selected designs. *After visual verification of the best energy combinations, the above mutations were chosen to redesign TEV proteases for in vitro assays.*

3.1.1.5 Other Key Positions and Combinations

The same *in silico* strategy was followed for the other two key positions of the substrate canonical sequence, P_6 (Glu302) and P_3 (Tyr305). In fact, some positions were calculated (see Table 3.4) and the rest of amino acids were partially screened *in silico* as well. These key positions, especially P_3 that is interacting with P_1 and TEV protease, are also very important to obtain a successful and very specific redesign. These combinations were not tested experimentally.

SUBSTRATE POSITION MUTATION	TOP RESIDUES IN TEV PROTEASE POSITION					DESIGNS
	K141	N171	T175	N176	Y178	
E302H	KRQEM	NADE	TNDEQY	NADESTQ	YF	1680
E302K	KRQEM	NADE	TNDEQY	NADESTQ	YF	1680
E302R	KE	NADET	TNDEQYA	NADESTQ	YF	840
		D148	N174	K220		
Y305F	-	DQMSRNE	NQEMLK	KRP	-	126
Y305H	-	DNQE	NDEQFWY	KRDENQA	-	196
Y305K	-	DNQE	DNQE	KRA	-	48
Y305N	-	DRQEHIL	NRQHILK	KRQHIL	-	294
Y305W	-	DRNQIK	NRDQEGHK	KRHLPA	-	288

Table 3.4 The other two key positions. *The best residues for each position of TEV protease interacting with the basic amino acids at the P_6 position are shown in blue. The best interactions with aromatic and polar amino acids at the P_3 position are shown in purple. The right column shows the number of designs to be calculated.*

3.1.2 In vitro Assays

3.1.2.1 TEV protease Production

The four designed proteases were obtained by site-directed mutagenesis as explained in Methods, and the corresponding proteins were expressed and purified. In Figure 3.6, a stained SDS-PAGE gel is shown, corresponding to the production and purification of the wild-type (wt) His-TEV protease. In all cases, the expression showed a certain amount of

the induced protein going into inclusion bodies. Nonetheless, the average amount of protein purified was enough to perform the experiments in all cases: wt TEV protease gave around 3 mg/ml final concentration, per liter of starting culture, and mutant TEV proteases gave around three times less effective concentration.

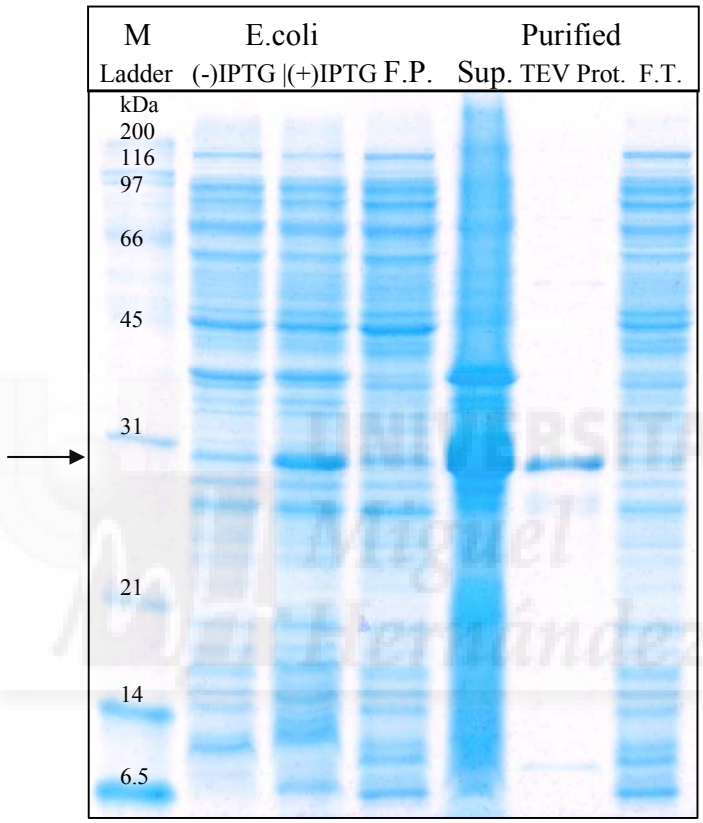


Figure 3.6 Expression and purification of wt TEV protease. *Lane: M = Standard Broad range ladder (Biorad); 1. E. coli Bl21(DE3)pLysS before the induction. 2. Proteins from E. coli Bl21(DE3)pLysS after the induction with IPTG. 3. E. coli Bl21(DE3)pLysS after lysis by French Press (F.P). 4. Supernatant after lysate clearance. 5. Purified TEV protease after the elution from the column. 6. Sample taken from the flow-through (F.T.) step before the elution; the arrow shows the 29 kDa TEV protease.*

3.1.2.2 Substrate-reporter production

The substrate-reporters were obtained by site-directed mutagenesis of the construct carrying the canonical sequence of the TEV protease target site expression vector. The corresponding substrate proteins were expressed and purified as described in Methods. In Figure 3.7 an SDS-PAGE gel is shown, corresponding to the production and purification

of "GST-canonical_substrate-GFP reporter" (see Fig.6.1). In all cases, the expression did not show the induced protein going into inclusion bodies. The average yield of wt substrate-reporter, after purification and quantification, was around 12 mg/ml per liter of starting culture. In the case of the mutated substrate-reporters, this yield was around half. All the purified proteins were diluted to 1 mg/ml to be used in the *in vitro* assays.

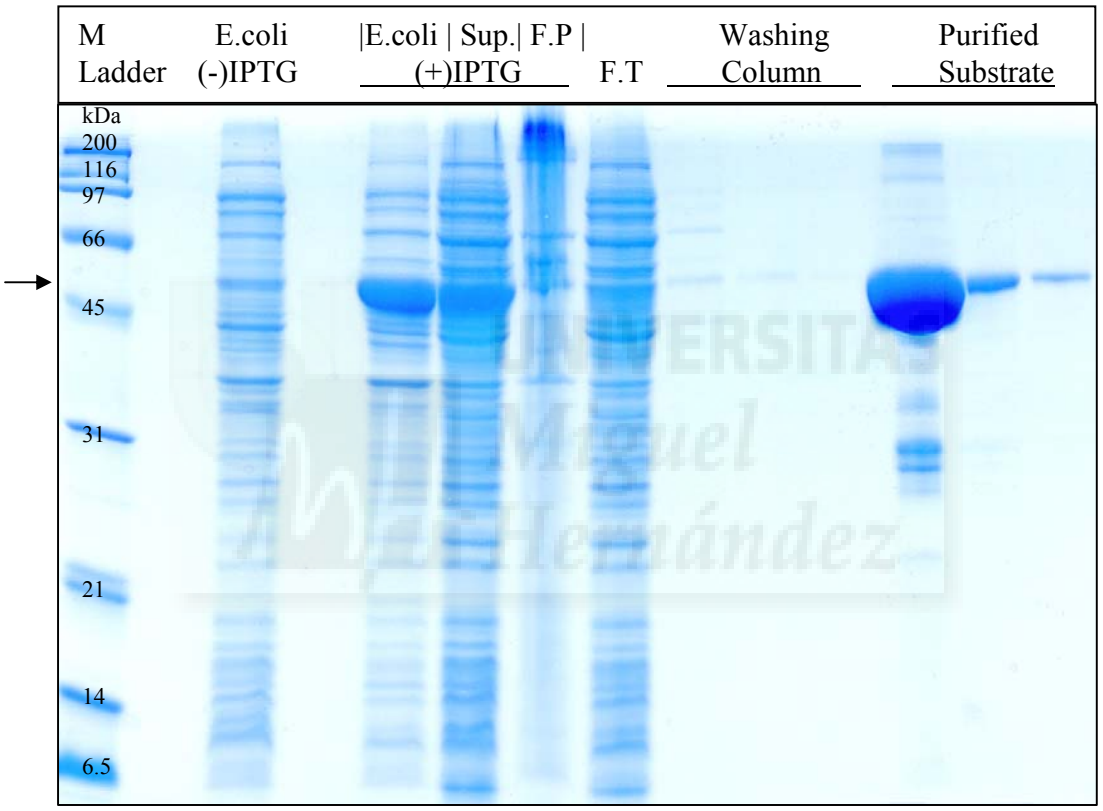


Figure 3.7 Expression and purification of Wt substrate-reporter. *Lanes: M = Standard Broad range ladder (Biorad); 1. Proteins from E.coli Bl21(DE3) before adding IPTG. 2. after the induction with IPTG. 3. Supernatant after lysis clearance. 4. After lysate. 5. Column GST flow-through (F.T). 6-8. Samples taken from the washing step before the elution. 9-11. Purified substrate-reporter after the elution; the arrow shows the 49 kDa substrate-reporter.*

3.1.2.3 Testing wt TEV protease

In order to demonstrate that the amino acid change in the cleavage pocket was indeed cleaving the desired target, the new designed TEV proteases were tested *in vitro*.

Hence, a digestion protocol was set up by using the wt TEV protease as the reference, and the redesigned TEV proteases were incubated with different substrate-reporters.

For the wt TEV protease, it has been reported that the standard reaction conditions for cleavage of the canonical target sequence were: 50 mM Tris-HCl, 0.5 mM EDTA and 1mM DTT, at pH 8.0, typically overnight (o/n), although cleavage happens within the first hours. TEV protease is maximally active at 34°C, but also works at room temperature (25°C). Notably, incubating at 4°C makes TEV protease only three-fold less active (Nunn et al., 2005).

Different ranges of enzyme:substrate (E:S) concentration ratios have also been reported (Kapust et al., 2002). Therefore a wide set of conditions were assayed, for wt TEV protease cleaving the Q307Q substrate-reporter and the Q307D substrate-reporter. These assays allowed us to find the optimal conditions to measure the activity and the specificity of wt TEV. Using the standard reaction buffer at 25ºC, several incubation times were tested (0, 1h, 2h, 3h, o/n) for four concentration ratios (1:12.5, 1:25, 1:50 and 1:100) (see Fig.3.8). Consequently, the ratio 1:50 and room temperature were chosen as the optimal conditions to perform all the cleavage assays. With regards to timing, the results showed that wt TEV protease cleaves the canonical substrate-reporter almost completely in the first three hours of reaction at the highest E:S assayed (see Fig.3.8, bottom left panel), whereas at higher E:S ratios (see Fig.3.8, left panel), a full digestion is only obtained after an overnight digestion. The right panel of Fig 3.8 shows the cross-reaction of wt TEV protease and mutant substrate Q307D.

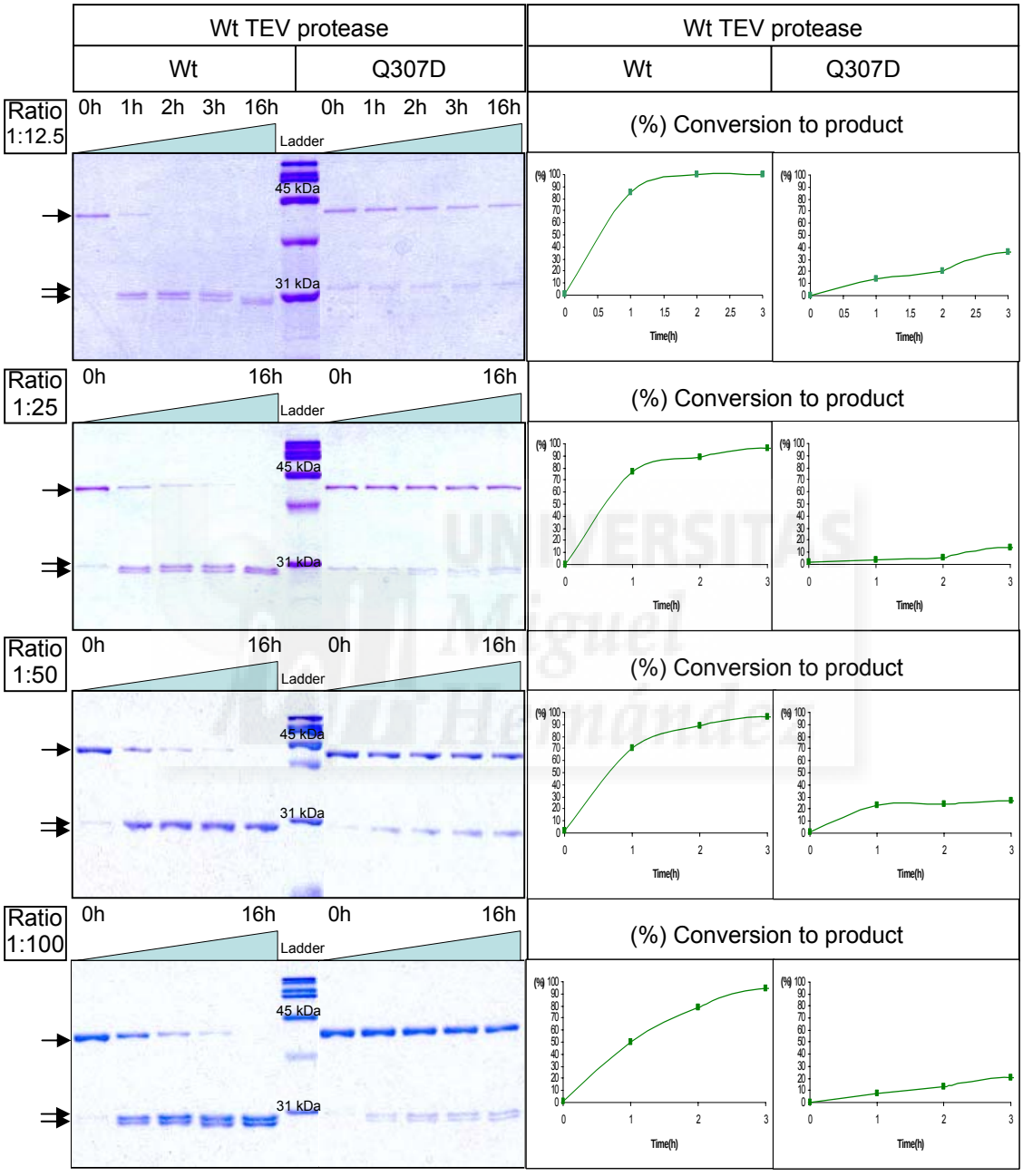


Figure 3.8 Enzyme:Substrate ratio assays. *Left panels: Lanes show the cleavage times 0h, 1h, 2h, 3h and 16h, of wt TEV protease plus wt substrate-reporter (on the left) and the wt TEV protease plus Q307D substrate-reporter. The single arrows mark the substrate-reporter, double arrows mark the products. Right panels: Quantification plots show conversion percentage of substrate to product.*

Although it was expected to have null activity against the mutated substrate, wt TEV protease showed a degree of cross-reactivity against the substrate Q307D, albeit a markedly decreased one. Figure 3.8 (right panel) shows the quantification of the SDS-PAGE gels (see 6.5.3), where the percentage of conversion to product is plotted versus digestion time. In all cases, but markedly in the lowest E:S ratios, the wt ligand showed an exponential decay that tended towards zero detection after a few hours of reaction. By contrast, the mutant Q307D was only barely digested (horizontal line) after several hours.

3.1.2.4 Optimizing the Cleavage Assays of wt TEV protease

The conditions of the cleavage assays were further optimized since it has been reported in the literature that TEV protease can be active over a wide range of pH and salt concentrations (Nallamsetty et al., 2004). Thus, the wt TEV protease assay was repeated against the wt and mutant Q307D substrate-reporters, varying the reaction conditions as follows: a set of reactions ranging pH between 6 and 9 were made, as well as some tests changing the concentration of NaCl while fixing the pH (see Fig.3.9). The other cleavage conditions used were: room temperature, overnight incubation, and the 1:50 enzyme:substrate ratio.

The analysis and quantification of the SDS-page electrophoresis (see Fig.3.9 and 3.8) showed a prominent decrease of the wt TEV protease activity at the highest pH assayed. The optimal conditions were around pH 7 or 8.

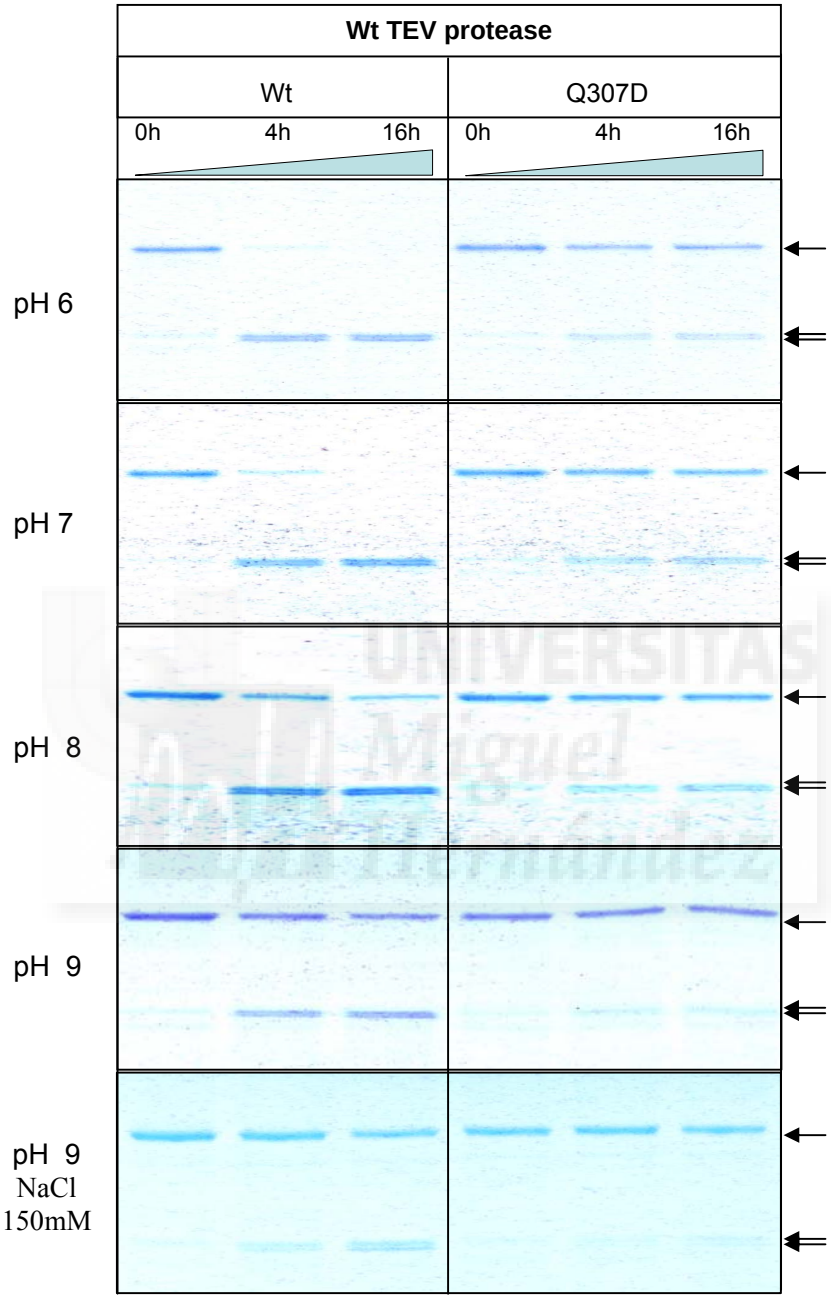


Figure 3.9 Effect of pH and NaCl on cleavage of Wt TEV protease. *The lanes on the left show the wt TEV protease, plus wt substrate-reporter, under low (50mM) or high (150mM) NaCl concentration, and different pHs, at time 0h, 4h and 16h. The lanes on the right show the wt TEV protease with the Q307D substrate-reporter. Single arrows mark the Substrate-reporter, double arrows mark the cleavage products.*

3.1.3 Comparison of the Activities of the wt and Designed TEV proteases

The designed TEV protease mutants (ASRK, ATRK, MSLK, MTLK) were assayed against wt and Q307D substrates to check whether the computational design was able to modify the specificity of the proteases. Figure 3.10 shows the overnight digestions of mutant TEV proteases (panel B: ASRK, panel C: MSRK, D: MTLK and E: ATRK) against wt and Q307D substrates. The digestion conditions were: room temperature, overnight incubation, and the 1:50 Enzyme:Substrate ration, with reaction buffer containing 150mM NaCl and pH 8. Unfortunately, most of the designs were completely inactive. Only the design ASRK showed activity against the substrate Q307D and also high cross-reactivity (Fig 3.10, panel B, lanes for the wt, and lanes Q307D).

3.1.3.1 Optimizing the cleavage of mutant designs

The low activity found in the mutant TEV ASRK against Q307D is probably due to activity loss after redesign, or changes in the optimal activity conditions of the enzyme. In order to investigate if the loss activity was due to the digestion conditions, the ASRK mutant activity was checked at different pHs, temperatures or E:S ratio, in the presence of the wt and Q307D substrates.

Figure 3.11 shows that incubating at pH 8 slightly increases the activity of mutant ASRK against substrate Q307D, and slightly decreases the cross-reactivity against wt substrate. Thus high salt concentration was used (150mM NaCl) and pH 8 in the subsequent digestions. The other conditions were as before: room temperature and 1:50 enzyme/substrate ratio.

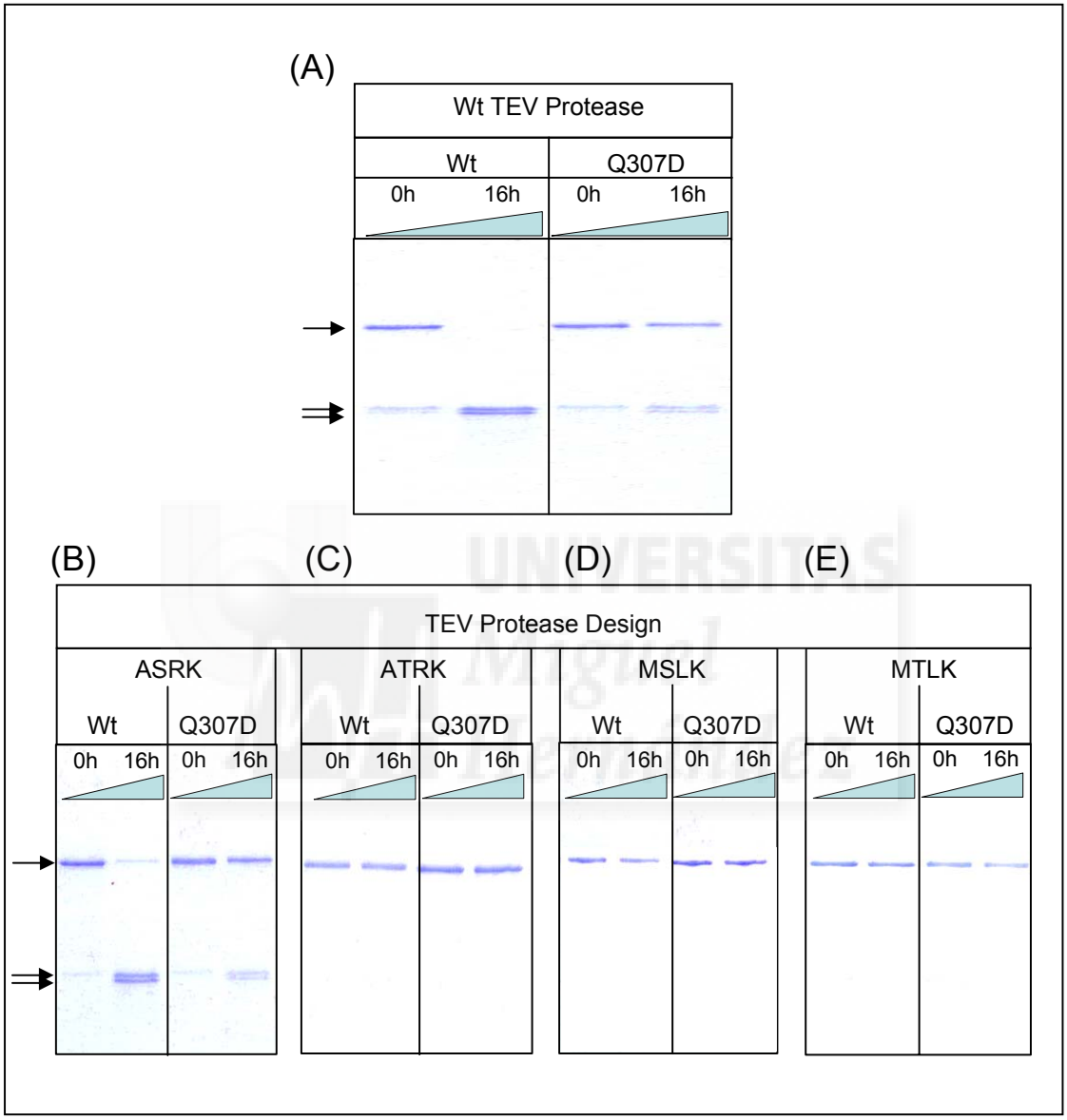


Figure 3.10 Cleavage assays of designed TEV proteases. *(A) wt TEV protease plus wt substrate-reporter (left) and Q307D substrate-reporter (right) after 0h and overnight reaction (16h). (B) Design 1 TEV protease: T146A, D148S, H167R, N174K. (C) Design 2 TEV protease: T146A, D148T, H167R, N174K. (D) Design 3 TEV protease: T146M, D148S, H167L, N174K. (E) Design 4 TEV protease: T146M, D148T, H167L, N174K. The arrowheads mark the cleavage product of the successful designed TEV protease. Single arrows mark the Substrate-reporter, double arrows mark the products.*

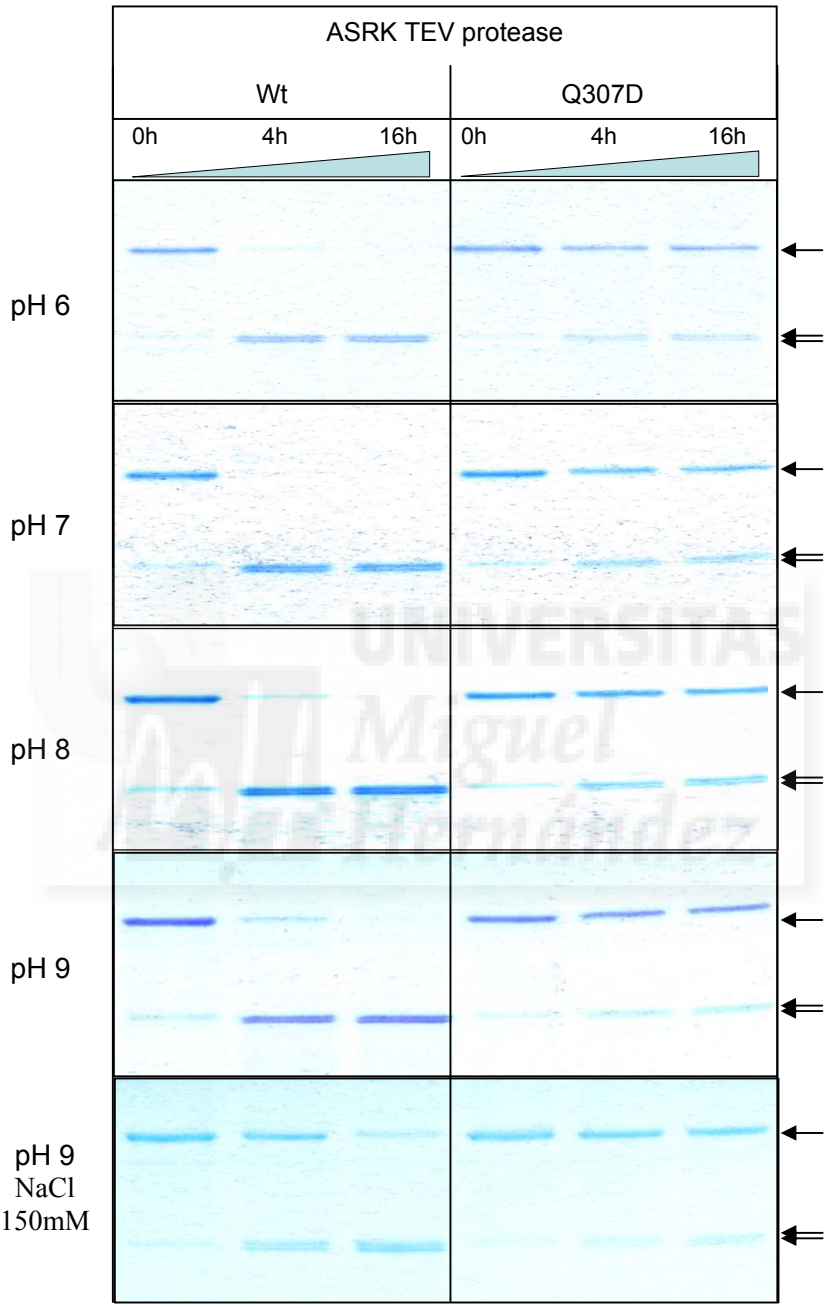


Figure 3.11 Effect of pH and NaCl on cleavage of ASRK TEV protease. *The lanes on the left show the ASRK TEV protease plus wt substrate-reporter cleavage assay under low (50mM) or high (150mM) NaCl concentrations, and at different pH, at times 0h, 4h and o/n. The lanes on the right show the ASRK TEV protease with the Q307D substrate-reporter. Single arrows mark the substrate-reporter, double arrows mark the products.*

3.1.3.2 Comparative kinetics of wt TEV protease and the ASRK design

The compared kinetic assay of wt TEV protease and the mutant ASRK, showed relatively low activity and low specificity of the mutant enzyme after 12 h of reaction (see Fig.3.12). For comparative assays, the same conditions were used as before: temperature of 22ºC, E:S ratio 1:50, 150mM NaCl and pH 8 (the optimal pH for ASRK mutant). The ASRK TEV protease cleaves 25% of the Q307D substrate, which is similar to the wt protease cleaving the same substrate.

In summary, while other mutants were completely unsuccessful, the ASRK design was an intermediate situation, where the activity against wt substrate was decreased to around 75%, while the activity against substrate Q307D was slightly increased. Therefore our protein design exercise was only partially successful and further mutants would have to be screened to attempt improving the ASRK design.

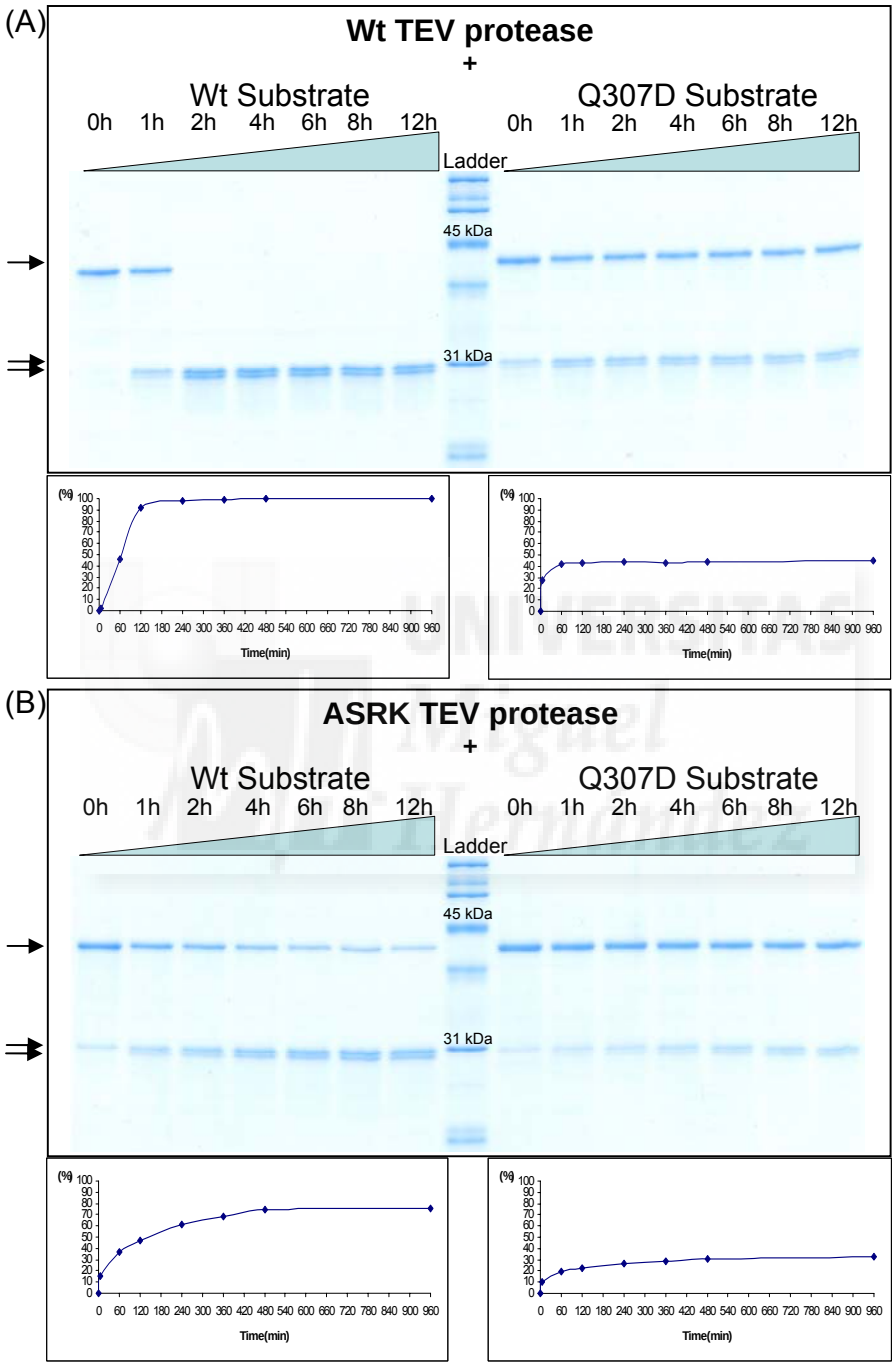


Figure 3.12 Kinetic assays of wt TEV protease and the ASRK TEV protease. *SDS-PAGE gels and graphical representation of the percentage of substrate converted to product, plotted against time. (A) wt TEV protease (B) ASRK TEV protease against the wt substrate-reporter and Q307D substrate-reporter. The best conditions for ASRK TEV protease were used; Enzyme:Substrate 1:50 ratio, 150mM NaCl, pH 8.0 and 22ºC.*

3.2 Meganuclease Interface Redesign

3.2.1 Computer-Aided Protein Design

To design a heterodimeric interface for I-CreI, the X-ray structure of the homodimer bound to its cognate DNA target sequence was used. This structure was determined at 2.05 Å resolution (PDB:1g9y)(Chevalier et al., 2001b). The aim was to facilitate heterodimerisation and at the same time to prevent the formation of homodimers, or at least make them thermodynamically unstable.

The starting point was the visual inspection of the interaction interface. A large part of the dimerization interface of the homodimer is composed of two α-helices (Lys7 to Gly19 in both monomers), arranged in a coiled-coil fashion. The two helices are very close to each other, packing in the centre mainly through the backbone, making them unsuitable for re-design. The amino acids below these helices (Asp20 and onward), are contacting the DNA and are thus responsible for both the activity (active site) and specificity (DNA recognition site) of the endonucleases (see Fig.3.13 panel A). These functions alone prevent any of these residues to be modified in the design process.

Because of the importance of these areas, there were few possibilities to enforce the heterodimerisation. After careful examination of the structure, three patches were identified, involved in the interactions interface that could be disturbed and changed in the dimers, without impairing their binding capacity or their enzymatic activity (see Fig.3.13 panel B).

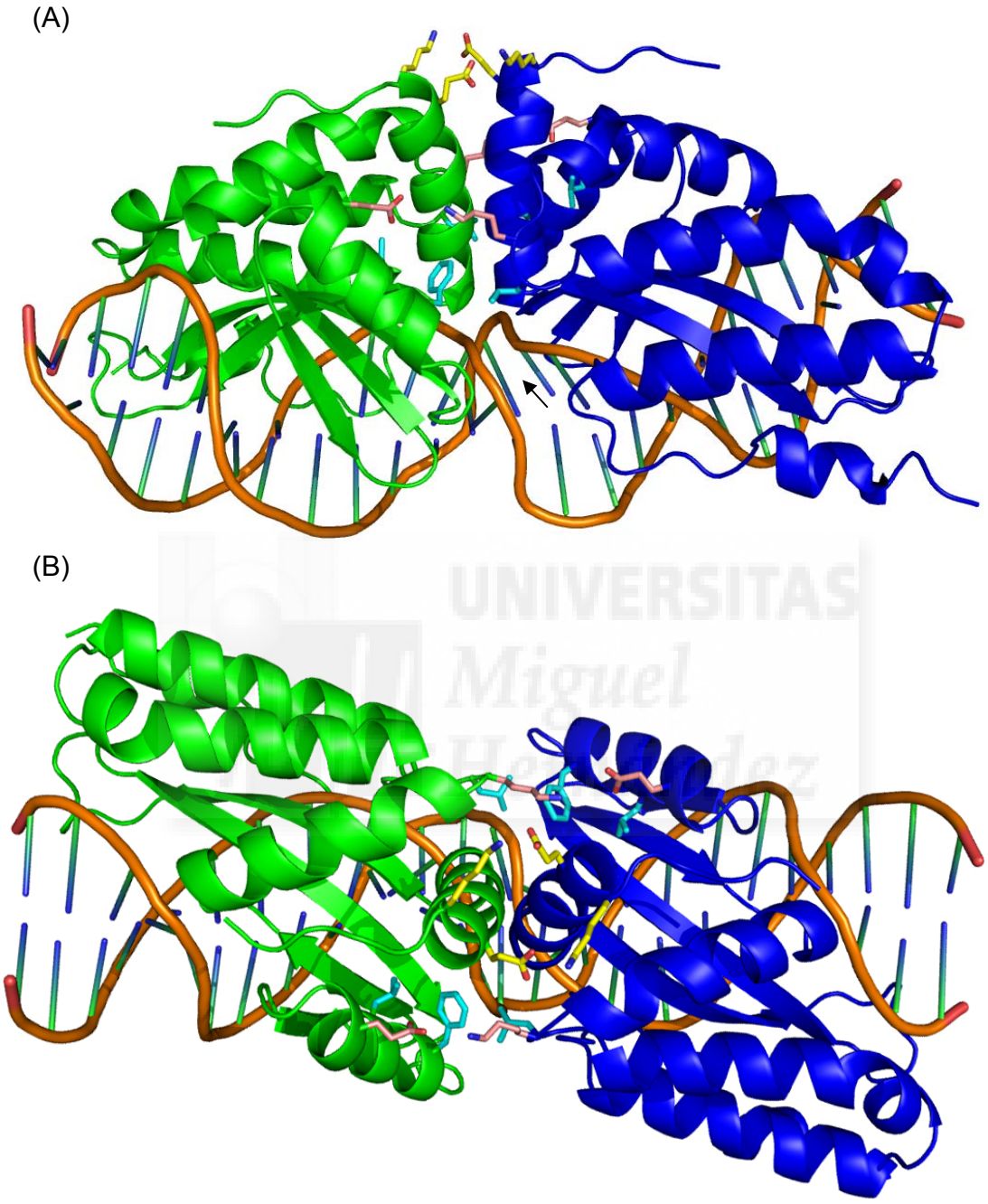


Figure 3.13 Structure of the complex of meganuclease I-CreI (PDB: 1G9Y) binding to its target DNA site. *(A) Side view and (B) Top view. Cartoon representation of monomer A (for clarity in green and blue) forming the wt homodimer. Amino acids (in sticks) show the three modifiable interaction patches between the two monomers on the homodimer. An arrow shows the cleavage site.*

3.2.1.1 Redesigning Patches

The most apparent interaction region is located above the two helices (see Fig.3.14 left panels), where Lys7 and Glu8 in one monomer establish favorable electrostatic interactions with the corresponding residues in the other monomer. In order to keep this interaction in the heterodimer, and at the same time impair monomer formation, it was decide to replace them with two arginines in one monomer (named monomer A hereafter) and two glutamates in the other (called monomer B). Thus, AA and BB homodimers would undergo an electrostatic repulsion whereas AB heterodimer formation would be electrostatically favorable (see Fig.3.14 right panels).

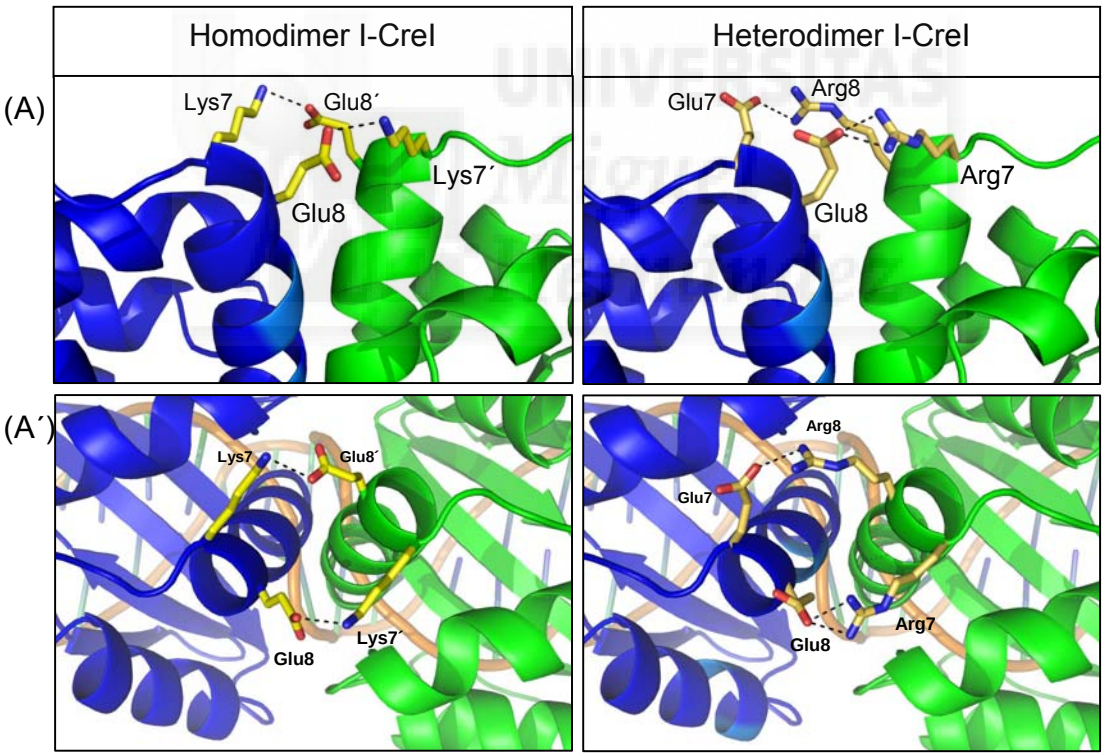


Figure 3.14 Detail of the first common modifiable interaction patch among monomers of meganuclease I-CreI. *(A) Side view and (A´) Top view. Left panels show the cartoon representation of wt monomers A (K7, E8 and K7´, E8´) forming the wt homodimer. Right panels show the cartoon representation of the mutant Ax monomers (R7, R8) and mutant Bx monomers (E7, E8) forming the obligatory heterodimer.*

The second patch was chosen with the same idea of creating small electrostatic imbalances for homodimers, relative to heterodimers. This patch is positioned on each side of the coiled-coil; again a double cluster of charged residues is made by the Lys96 and the Glu61 of each monomer (see Fig.3.15 left panels). To re-enforce the electrostatic effects of the first mutation site, the second site was mutated with two arginines in monomer A, and two glutamates in monomer B, thus making a charged triangle in each monomer (positive in A, negative in B) (see Fig.3.15 right panels).

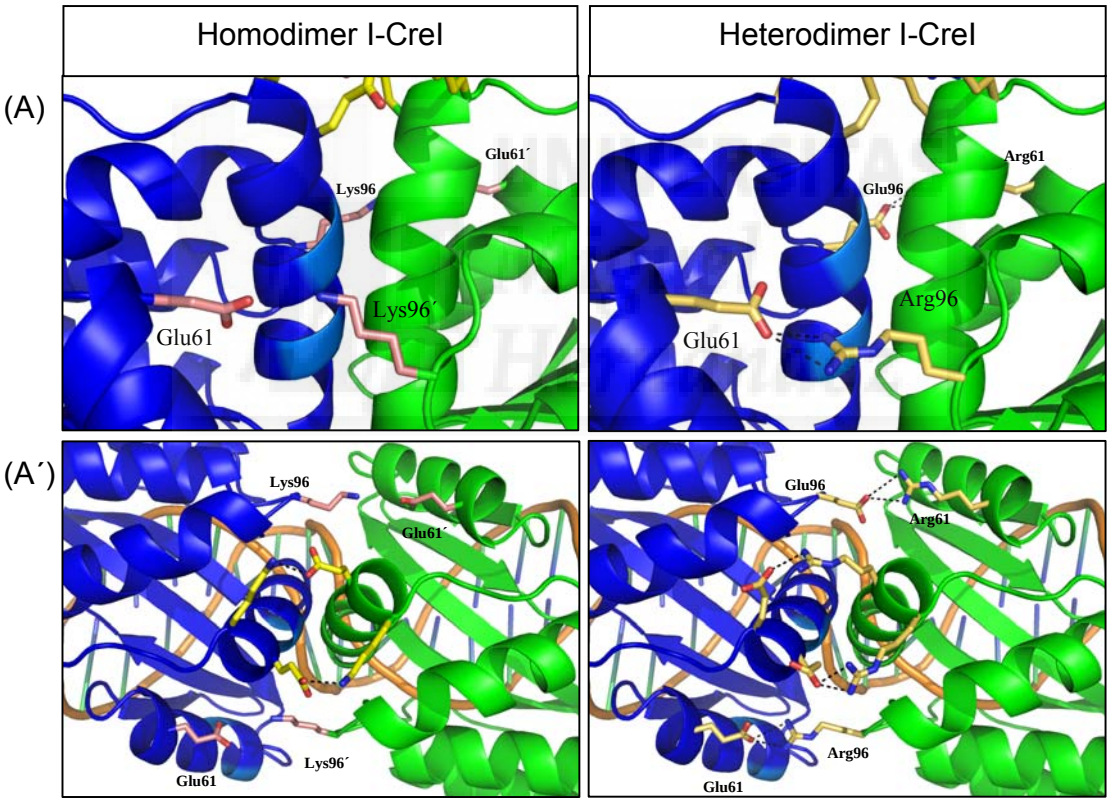


Figure 3.15 Detail of the second common modifiable interaction patch among monomers of meganuclease I-CreI. *(A) Side view and (A´) Top view. Left panels show the cartoon representation of wt monomers A (E61, K96 and E61´, K96´) forming the wt homodimer. Right panels show the cartoon representation of the mutant Ax monomers (E61, E96) and mutant Bx monomers (R61, R96) forming the obligate heterodimer.*

The third region of interest is the region around the middle of the two helices involved in the interaction surface and is mainly composed of hydrophobic interactions and hydrogen bonds, making a kind of minicore. As the Hydrogen-bond network is quite intricate and extends all the way to the active site, only one hydrophobic patch was perturbed, involving residues Tyr12, Phe16, Val45, Trp53, Phe54, Leu55 and Leu58 of one monomer, interacting with residue Leu97 of the other monomer (the latter acting like a cap closing the hydrophobic pocket) (see Fig.3.16).

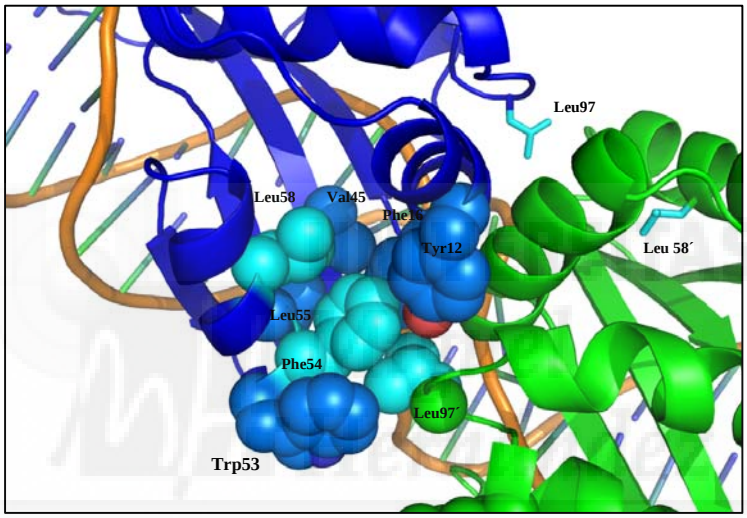


Figure 3.16 Detail of the lateral "minicore" patch among monomers of meganuclease I-CreI. *Amino acids implicated on one lateral "minicore" of the monomers forming the wt homodimer, are represented in Balls. The residues mutated in the monomers, to allow the formation of the obligate heterodimer, are in light blue.*

In order to introduce strong Van der Waals' Clashes in the homodimers, without disturbing the hydrophobic interactions in the heterodimers, these pockets were redesigned (i.e. without creating cavities and steric clashes). For this, it was decided to introduce bulky residues in monomer A (respectively Phe or Trp for position 54 and Phe for position 97 and small residues in monomer B; Gly and Leu, respectively). A Glycine was introduced at position 97 to give more space to position 54, for constructs where the latter position was mutated to Tryptophan.

Thus, two types of monomer A were defined (A1 and A2), depending of the nature of the amino acid at position 54 (Phe or Trp, respectively) and two types of corresponding monomer B (B3 and B4), the later differing by a mutation in Glycine at position 97 to accommodate with the Trp mutation of monomer A2. Finally, Leu58 was mutated to methionine in monomers B, to prevent any cavity formation in the heterodimer, due to the introduction of the small side chains (see Table 3.5).

	Wt MONOMER	DESIGNED MONOMERS			
PATCHES	**RESIDUES**	**A1**	**A2**	**B3**	**B4**
Top	**Lys 7**	Arg	Arg	Glu	Glu
	Glu 8	Arg	Arg	Glu	Glu
Middle	**Glu 61**	Arg	Arg	Glu	Glu
	Lys 96	Arg	Arg	Glu	Glu
Minicore	**Phe 54**	Phe	Trp	Gly	Gly
	Leu 58	Leu	Leu	Met	Met
	Leu 97	Phe	Phe	Leu	Gly

Table 3.5 List of the proposed monomers, patches involved in the redesign and the suggested mutagenesis to get obligate heterodimers. *Top and middle patches are implicated in the electrostatic interaction between monomers, making one part of the dimer positively charged (monomer A) and the other part negatively charged (monomer B). The minicore patches are implicated in the formation of hydrophobic packing surfaces between both monomers.*

3.2.1.2 Energy Analysis

It was expected that AA homodimers thus develop huge electrostatic and steric hindrance, preventing their formation, while BB homodimers will suffer also the charge repulsion and contain big cavities making them unstable. By contrast, the minicore of AB heterodimers should be filled efficiently by these compatible amino acids and must reinforce the upper patches with complementary charges.

The different mutations were made with FoldX to model all homodimers (A1:A1, A2:A2, B3:B3 and B4:B4) and heterodimers (A1:B3, A2:B3 and A2:B4) and to get the different interaction energies. The energies were compared with the wild type energies to see whether the binding was improved in the new designs (see Table 3.6).

Dimers	ΔΔG between mutants and wild type (kcal/mol)
A2_B3	0.13
A1_B3	0.22
A2_B4	3.20
B3_B3	7.39
A1_A1	8.30
A2_A2	8.53
B4_B4	11.96

Table 3.6 FoldX calculated interaction energies (kcal/mol) between wild type and designed homodimers and heterodimers. *Differences in interaction energies below 3.5 kcal/mol are highlight in green. The best binding energy (A2_B3) corresponds with the best "in vitro" result.*

Of all the heterodimers, two constructions, A1:B3 and A2:B3, presented a computed interaction energy close to the wild-type homodimer (Table 3.6). The last construction, A2:B4, presented a significant decrease in interaction energy compared to

the wild-type homodimer but was nonetheless significantly higher than the homodimers. Conversely, A1:A1, A2:A2, B3:B3 and B4:B4 homodimers were all much destabilized and thus these species were expected to remain monomeric.

3.2.2 Optimizing Conditions for Specific DNA Cleavage.

To verify that it was possible to design a specific heterodimer correctly, two meganuclease variants that recognize different DNA sequences were employed (see 1.4.3.2). These I-CreI variants both harbor an Asp75 to Asn mutation that decreases energetic strains caused by the replacement of the basic residues Arg68 and Arg70; these arginines normally satisfy the hydrogen-acceptor potential of the buried Asp75 in the I-CreI structure.

Hereafter, the meganuclease denoted as "KTG" differs from the wt at positions Q44, R68 and R70 (implicated in the specific binding with the DNA target) and recognizes the bases CCT at positions 4, 5 and 6 of the DNA target. The other meganuclease is called "QAN", differs from the wt at the same positions, and recognizes the bases GTT at positions 4, 5 and 6 of the DNA target.

Throughout this thesis, the target DNA sequences are denoted by a 6-base code. The first three bases corresponding to positions -5, -4, and -3 of the pseudo-palindromic target sequence, and the second three (3, 4 and 5) to the same positions in the complementary DNA strain, with the two triplets separated by a slash (/); see example below (for more details see 6.3.2.3).

-5 -4 -3 / 3 4 5

Thus the target of the KTG enzyme is CCT/AGG, that for the QAN enzyme the target is GTT/AAC, and the mixed sequence target for the heterodimer KTG-QAN is denoted as GTT/AGG.

3.2.2.1 Cleavage Specificity and Ionic Strength

For the wt meganuclease I-CreI, it has been reported that the ideal conditions for digestion of its target DNA are: 10-20mM Tris–HCl (pH 7.0-9.0) with 10 mM $MgCl_2$, at 37ºC and the enzyme is reportedly inhibited above 25mM NaCl ionic strength (Wang et al., 1997). When using these conditions with the KTG and QAN enzymes, suboptimal specificity was actually found (see Fig.3.17). In fact, at low ionic strength (≤50mM NaCl) both enzymes digest not only their target DNA sequence but also the mixed DNA target. This suggests that strong binding of only one of the monomers to the DNA is enough to allow digestion.

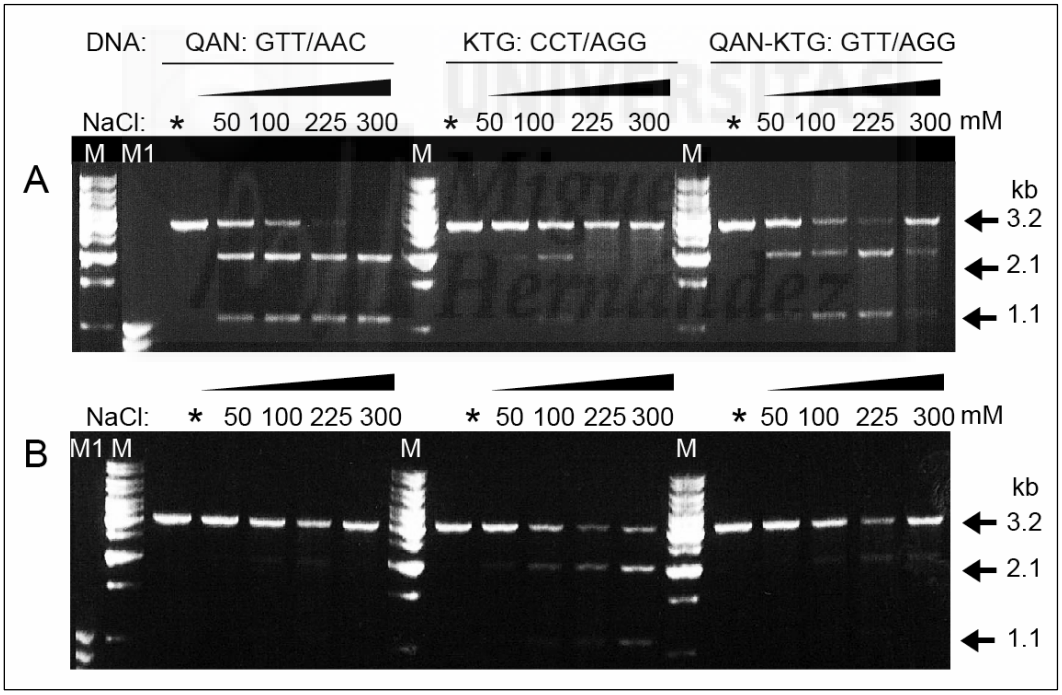


Figure 3.17 High salt concentration increases the cleavage specificity. *(A) QAN or (B) KTG protein was incubated with either the QAN homodimer site (GTT/AAC), the KTG homodimer DNA site (CCT/AGG), or a hybrid site QAN/KTG site (GTT/AGG), varying the concentration of NaCl between 50 and 300 mM. Arrows indicate the uncut target DNA (3.2 kb) or the two bands resulting from digestion (1.1 and 2.1 kb). An asterisk (*) marks control lanes with DNA alone.1 kb and 100bp ladders (Fermentas) are marked by M and M1, respectively.*

Increasing ionic strength both improves the activity of the enzymes towards their targets and reduces the digestion of the mixed template: at around 225 mM NaCl, almost perfect specificity and good activity were found. This behavior could be explained by the ionic strength decreasing the affinity for DNA (thus preventing binding if only one monomer establishes specific interactions in the dimer), while also increasing enzymatic activity. As a result of these tests, the following optimal buffer for digestion of the meganuclease designs was selected: 25mM HEPES, 5 % Glycerol, 10mM $MgCl_2$ and 225mM NaCl, pH 8 (see 6.5.4).

3.2.3 Expression and Characterization of the Designed Mutants

The designed mutants A1, A2, B3 and B4 were obtained by site-directed mutagenesis of the original KTG and QAN enzyme expression vectors, and the corresponding proteins expressed and purified (see chapter 6). Not every combination of possible variants was tested, but rather a representative selection: QAN-A1, KTG-B3, KTG-A2, QAN-B3 and QAN-B4. These were designed to give coverage of all the designed heterodimer interactions A1:B3, A2:B3 and A2:B4, resulting in the heterodimers QAN-A1:KTG-B3, KTG-A2:QAN-B3 and KTG-A2:QAN-B4.

Whereas the wild-type KTG and QAN enzymes' protein yield, after purification, contains the majority of protein in the soluble fraction, the opposite happened in the case of the designed enzymes: the majority of the expressed proteins remained in inclusion bodies in the pellet, only a small fraction could be purified, and even this was contaminated by other proteins. This was a first indication that the designed variants cannot homodimerize and thus become unstable and aggregate when expressed individually (see Fig.3.18).

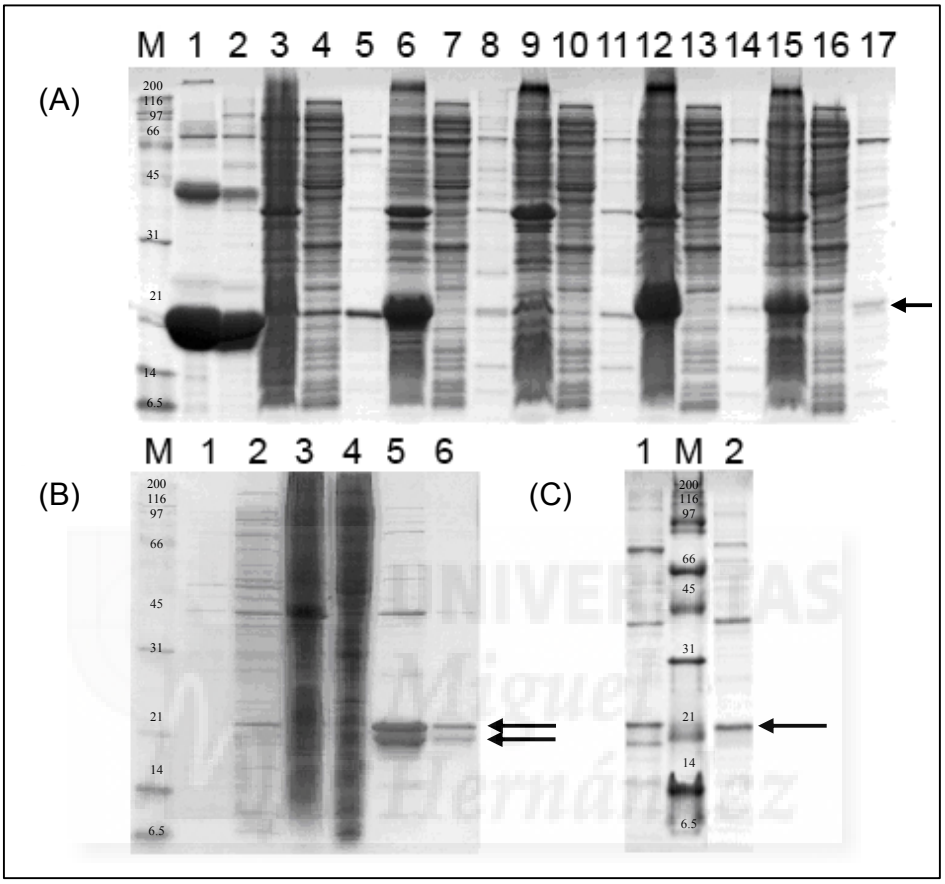


Figure 3.18 Expression and purification of the designed meganucleases. *The target protein bands are marked by arrows. (A) Wild-type homodimers and mutant monomers. Lanes: M = Standard Broad range markers (Biorad); 1. Purified QANwt; 2. Purified KTGwt; 3. Pellet QAN-A1; 4. Supernatant QAN-A1; 5. Purified QAN-A1; 6. Pellet KTG-B3; 7. Supernatant KTG-B3; 8. Purified KTG-B3; 9. Pellet KTG-A2; 10. Supernatant KTG-A2; 11. Purified KTG-A2; 12. Pellet QAN-B3; 13. Supernatant QAN-B3; 14. Purified QAN-B3; 15. Pellet QAN-B4; 16. Supernatant QAN-B4; 17. Purified QAN-B4. (B) Co-expression and purification of KTG-A2/QAN-B3. Lanes: 1. Uninduced; 2. Induced; 3. Pellet; 4. Supernatant; 5. Purification before dialysis; 6. Purified after dialysis. (C) Co-expression of the other two designs: Lane 1. Two bands are visible, corresponding to the heterodimer QAN-A1/KTG-B3; Lane 2. Only one band is visible, indicating that QAN-B4/KTG-A2 does not make a heterodimer.*

3.2.3.1 Testing the Activity of Mutants

Following purification, the activity of the A1, A2, B3 and B4 enzymes on the three DNA targets was tested (see Fig.3.19) at low and high ionic strength (50 mM or 225 mM NaCl respectively). At low salt concentrations, only some specific DNA digestion activity for QAN-A1 was detected; for the other enzymes no specific cleavage could be detected. Moreover, at high ionic strength the two expected DNA bands could not be detected either, although the amount of DNA decreased upon incubation with the enzymes, in some cases.

These results were marred by the low yield and quality of the protein obtained when the non-homodimerising monomer designs were expressed individually; even with a large 6-litre volume of bacteria yielding inadequate protein (between 0.5-1.5 mg/ml for the designed monomers compared with 30 mg/ml for wild-type dimerising monomers).

3.2.3.2 Testing Oligomerization

To check the oligomeric status of the purified designed enzymes, their size profiles were measured by analytical ultracentrifugation (see Fig.3.21 and 3.22). In the case of individually-expressed A1, A2, B3 and B4 proteins, the expected monomeric enzyme was observed. However, higher molecular weight aggregates were also seen; including trimers and tetramers (see Fig.3.19 B); for clarity only KTG-A2 and QAN-B3 are shown, although similar results were obtained with the other designs). Therefore the designed enzymes were indeed unable to homodimerize, and this may have affected their stability and aggregation properties during purification.

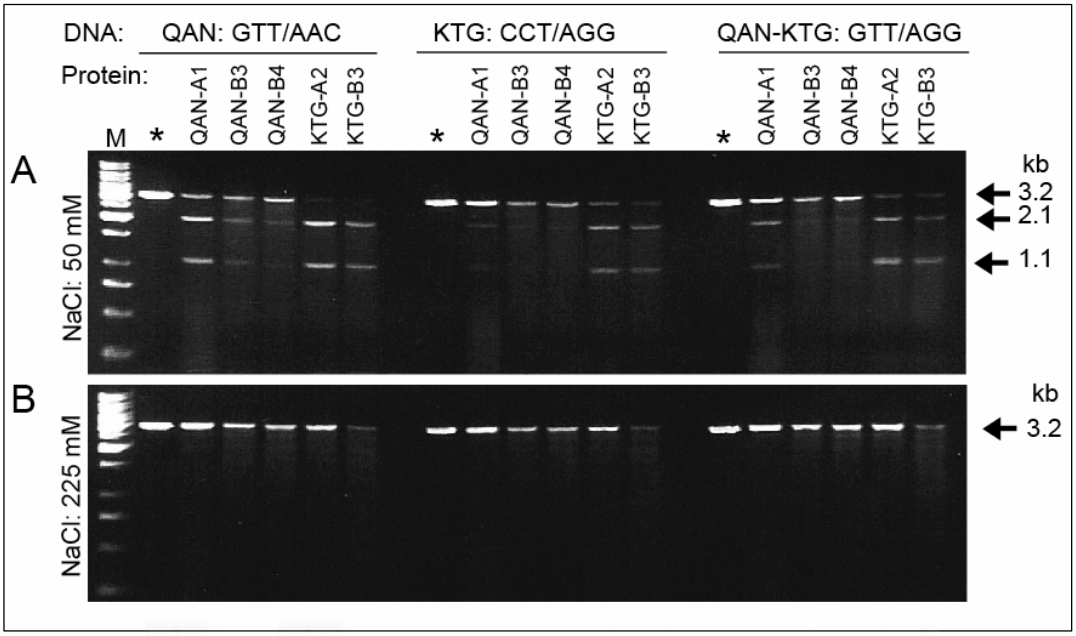


Figure 3.19 Non-specific DNA cleavage and non-cleavage by singly-expressed designed meganuclease monomer variants under different salt conditions. *3.75 µM of each purified protein was incubated with 34 nM of purified plasmid (pre-linearized with XmnI), containing either the QAN homodimer site (GTT/AAC), the KTG homodimer DNA site (CCT/AGG), or a hybrid site QAN/KTG site (Q-K: GTT/AGG). The concentration of NaCl was either (A) 50 mM or (B) 225 mM. Arrows indicate the uncut target DNA (3.2 kb) or the two bands resulting from digestion (1.1 and 2.1 kb). An asterisk (*) marks control lanes with DNA alone. 1 kb ladders (Fermentas) are marked by M.*

To investigate the potential for heterodimerisation, equimolar quantities of the individually purified designed enzymes (QAN-A1, QAN-B3, QAN-B4, KTG-A2 and KTG-B3) were mixed in all possible combinations. In the case of KTG-A2/QAN-B3 (the best heterodimer design), the appearance of a major species corresponding to the molecular weight of the dimer was shown, but this was not the only species formed (see Fig.3.20). For the pair QAN-A1/KTG-B3 and KTG-A2/QAN-B4, the appearance of new peaks of molecular mass between the monomer and dimer was revealed, and also a decrease of high molecular weight aggregates.

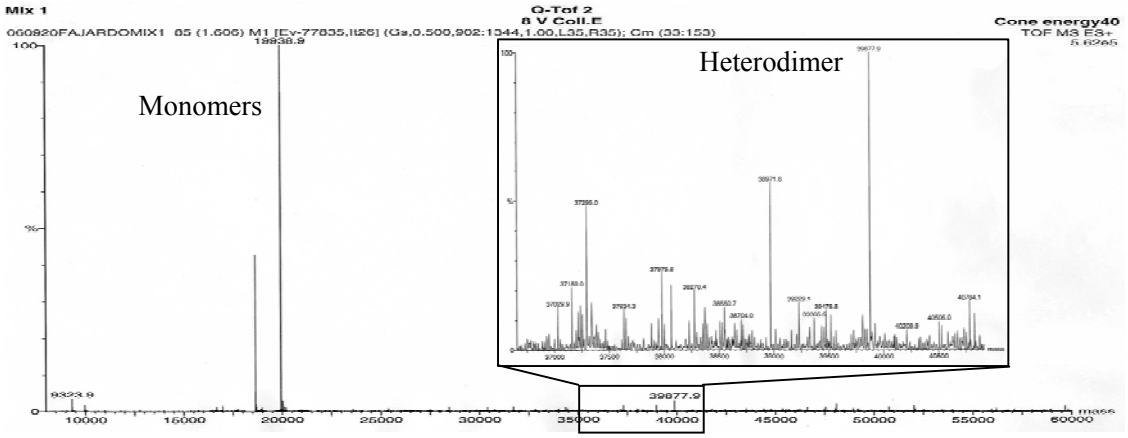


Figure 3.20 Mass Spectrometry analysis of the best heterodimer. *The KTG-A2 and QAN-B3 allow the obligate heterodimer formation. In the left part, appears the mass corresponding with the monomers and in the right, a zoom around 40 kDa shown the heterodimer formation.*

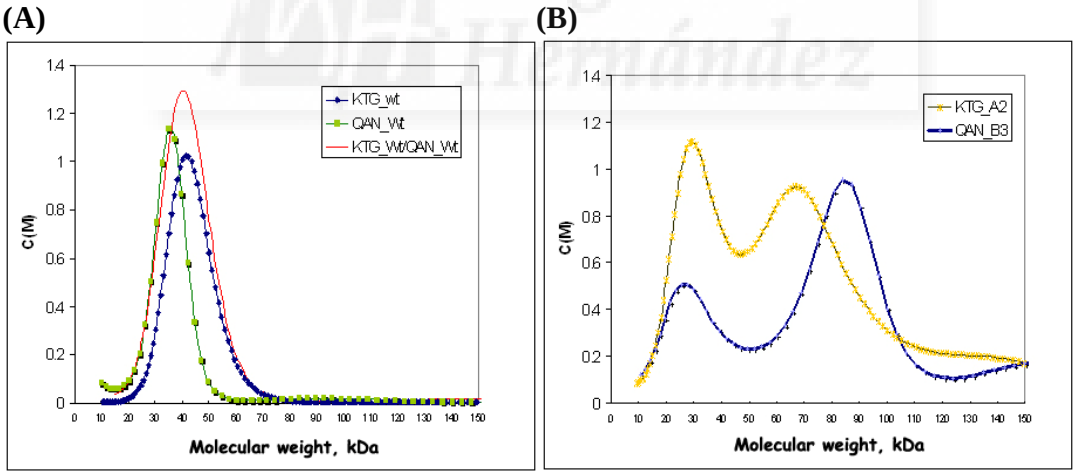


Figure 3.21 Analytical ultracentrifugation of the different meganucleases. *(A) The wild-type monomers form homodimers of about 40 kDa (KTG-wt; QAN-wt). (B) The designed non-homodimerising KTG-A2 and QAN-B3 form aggregates when expressed individually.*

For those combinations that should not produce a heterodimer, no significant changes in the behavior of the proteins were seen. Overall, these results indicated that the design might have been successful but that separate expression of heterodimerising monomers was not an effective strategy. Thus, it was decided to attempt co-expression within the bacterial cell.

3.2.3.3 Co-expression Assays

The above results suggested that the heterodimer designs might have been functioning, but that the expression of the monomeric enzymes resulted in strong aggregation and thus in partly inactive enzymes. To avoid this problem, the monomer gene expression cassettes were subcloned into complementary plasmids and co-transformed into bacterial cells, such that one monomer would be expressed (with a His-tag) from the original pET-series plasmid and that the partner monomer would be expressed (without a His-tag) from a compatible pCDFDuet-I vector (Novagen). Dual antibiotic selection ensured that each cell contained both plasmids.

Expression analysis of the co-expressed KTG-A2/QAN-B3 proteins showed that inclusion bodies were avoided, suggesting that the previous aggregation problem had been solved. SDS-page analysis of the purified enzyme subsequently revealed 2 bands with approximately the same amount of protein, suggesting that the heterodimer and not homodimer was purified (see Fig.3.18 Panel B). Mass spectroscopy directly confirmed the presence of the two proteins and of the heterodimeric complex (see Fig.3.20). Furthermore, an analytical ultracentrifugation of the purified proteins gave a clean single profile at the expected molecular weight for a dimer (see Fig.3.22 panel A).

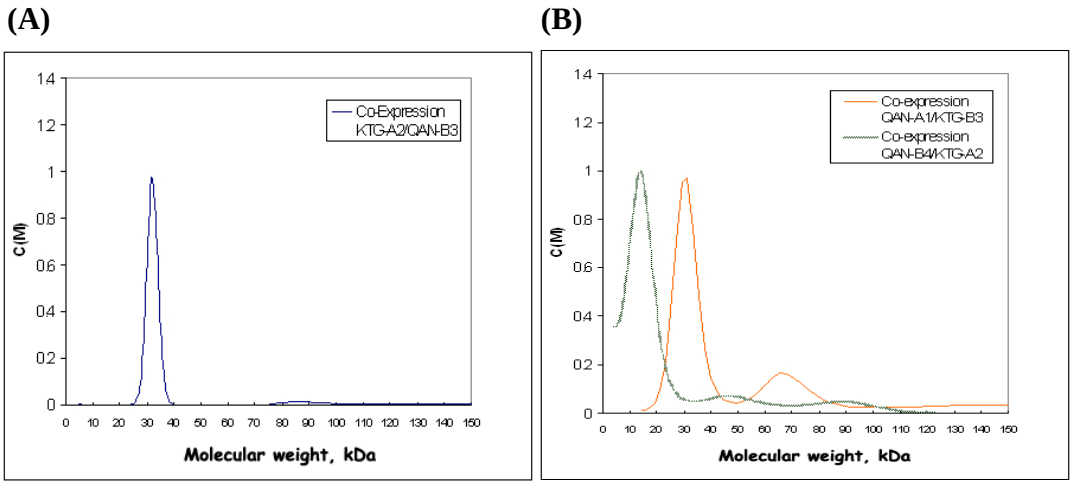


Figure 3.22 Analytical ultracentrifugation of the co-expressed meganucleases. *(A) The co-expressed KTG-A2 and QAN-B3 form a perfect heterodimer. (B) The co-expressed QAN- A1 and KTG-B3 also form a heterodimer, to an extent, while QAN-B4 and KTG- A2 do not.*

Cleavage assays of the various DNA targets with the purified co-expressed heterodimer designs were carried out (see Fig.3.23). Thereby, it was demonstrated that the KTG-A2/QAN-B3 design successfully gives a clear specific cleavage of the heterodimer DNA target (GTT/AGG), and not of the homodimeric targets (CCT/AGG and GTT/AAC). Thus the strategy of rational design, plus in silico screening with FoldX, has been successful in this case.

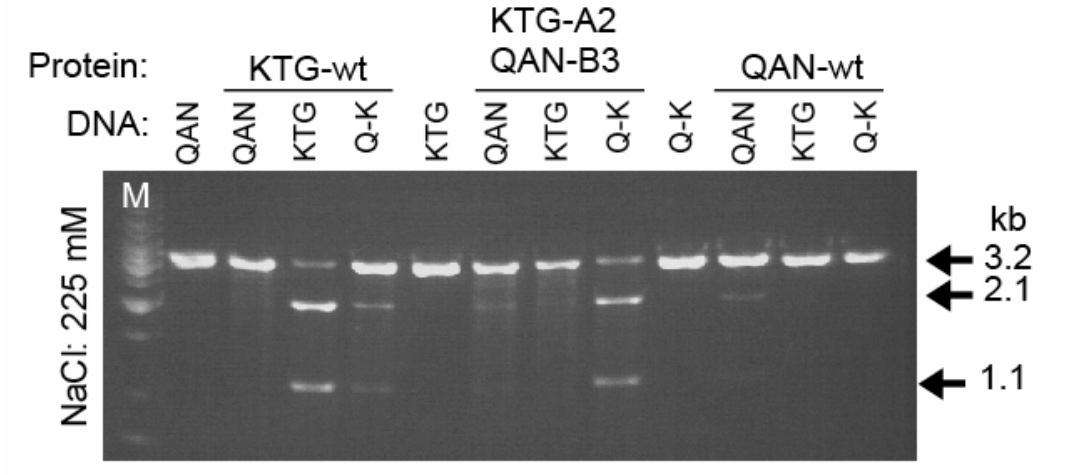


Figure 3.23 Specific DNA cleavage by co-expressed designed obligate heterodimer KTG-A2--QAN-B3 meganucleases. *3.75 µM of each purified protein was incubated with 34 nM of purified plasmid (pre-linearized with XmnI), containing either the QAN homodimer site (GTT/AAC), the KTG homodimer DNA site (CCT/AGG), or a hybrid site QAN/KTG site (Q-K: GTT/AGG). NaCl concentration was at 225 mM. Arrows indicate the uncut target DNA (3.2 kb) or the two bands resulting from digestion (1.1 and 2.1 kb). 1 kb ladders (Fermentas) are marked by M.*

The same experiments, when repeated with the co-expressed QAN-A1/KTG-B3 proteins showed mixed results; there were indeed 2 bands after purification, indicating heterodimer formation (see Fig.3.18 Panel C). However, one band was stronger than the other and while there was specific cleavage of the heterodimer, this was at a reduced level as compared to the KTG-A2/QAN-B3 combination. Analytical centrifugation showed formation of a dimer with a small proportion of aggregate. (see Fig.3.22 Panel B).

Finally the third co-expression combination, KTG-A2/QAN-B4, resulted in only one band being purified and a monomer detected by analytical centrifugation (see Fig.3.22 Panel B). Therefore this design failed to make a heterodimer, even when co-expressed (see Fig.3.18 Panel C).

Interestingly enough, the proportion of dimer and activity between KTG-A2/QAN-B3, QAN-A1/KTG-B3 and KTG-A2/QAN-B4 correlate perfectly well with the energies predicted by FoldX (Table 3.6). The best design KTG-A2/QAN-B3, in terms of predicted energy *in silico,* forms the best heterodimer *in vitro*, indicating that the redesign of the enzymes was successful.

In summary, the design KTG-A2/QAN-B3, was best in terms of predicted energy *in silico,* and also formed the best functional heterodimer *in vitro*. This indicates that the combination of rational design and FoldX verification of the enzymes was a successful strategy.

UNIVERSITAS
Miguel
Hernández

IV. DISCUSSION & PERSPECTIVES

One of the main interests of groups doing computational protein design is to produce proteins that bind to other proteins or DNA targets, with high specificity. Initial successful investigations have led researchers towards thinking about redesigning valuable proteins with customized features.

4.1 Computational Methods for Designing New Molecular Tools

This work shows how computational protein design can be successfully applied to obtain new enzymatic variants by using the methodology shown here. The protein design software FoldX has been already successfully used on some protein design projects, using similar strategies to those described here (van der Sloot et al., 2004; Reina et al., 2002; Kiel et al., 2004; Kiel et al., 2005; Kempkens et al., 2006; van der Sloot et al., 2006; Musi et al., 2006; Kolsch et al., 2007; Kolsch et al., 2007; Kolsch et al., 2007; Villanueva et al., 2003; Fernandez-Ballester et al., 2004).

In addition, combining computational and experimental methodology is a powerful approach in protein engineering; a preliminary *in silico* screening of the mutated structures helps to identify the most energetically favorable mutations and therefore decreases the number of variants to be produced and tested experimentally.

4.1.1 Challenges in Computational Protein Design

When using computational methods, some important critical factors affect the methods' accuracy. Therefore some limitations on their applicability must be considered, as for instance:

- Three-dimensional information, crystal structure refinement and the possibility of having crystallographic errors.

- Approximations of the force fields used, such as using a fixed backbone, rotamer search and the energetics.

- Final rational criteria to choose the most favourable designs to be produced.

Thus, the worse the resolution of the structure used as a scaffold, the lower the chances of a successful design. Even with high resolution structures, a preliminary repair should be done by slightly moving side chains that could have small van der Waals' clashes. In addition, an automatic 180 degrees search should be done for the side chains of asparagine, glutamine and histidine to ensure that they are in the correct conformation. In the case of His, the pH of the experiment is essential since it can either be charged or uncharged at near-physiological pH. Another critical point is to ensure that when mutating or moving side chains, the β carbon from the new residue is exactly placed on the crystallographic β carbon (except for proline and glycine). Small deviations on the canonical position when placing the β carbon could result in large deviations at the end of the side chain. Here, these steps are automatically carried out by FoldX.

The other common source of variability is the force field and the rotamer library used in the search. Most force fields, including molecular dynamics force fields, are empirical. In principle, the more accurate a force field is, the higher the reliability that should be expected when calculating energies. However, it will then be more sensitive towards crystallographic errors and to the denatured state heterogeneity. On the other hand, less accurate force fields will be less sensitive to these variables, but less precise in energy calculation. In any case, uncertainties in prediction will always be present because there is no such thing as a perfect force field (in the case of FoldX, the standard deviation in predicting point mutation changes in energy is around 0,7 kcal mol^{-1}).

Regarding rotamers, it is computationally very intensive to explore all possible conformations of a long side chain (such as lysine or arginine) at intervals of 1 degree (since in long side chains, changes of 1 degree in each dihedral angle can result in very large changes in the position of the last atom). Moreover, for mutating a position and moving all neighboring residues, it would require more than hundred full PhD time-periods just to explore all possible combinations. In spite of supercomputers and the fact that computational shortcuts are used, these do not guarantee escaping from conformational traps.

Ultimately, every design exercise is a compromise between computer time, exploration of sequence space and accuracy. Also the virtual structure output by the algorithm must be studied in detail and the "best data structures" must be examined to decide if they make biological sense or not.

4.1.2 In silico Design Using the FoldX Force Field

As mentioned above, FoldX already has been proven to be an accurate and efficient protein design software; the force field takes into consideration the main important issues in protein–protein interactions: solvation, van der Waals' interactions, hydrogen bonds, and the electrostatic and entropic terms for the backbone and side chains. In the case of protein complexes, the electrostatic contribution and the prediction of structural water molecules are also considered (see 1.2.4.1).

In addition, as an empirical force field, FoldX is programmed to allow fast and accurate estimations of free energy changes upon mutation in proteins (such as in the meganuclease interfaces) or protein complexes (such as TEV protease and its substrate binding pocket). This can be done with a similar accuracy to physical force fields for the prediction of free energy changes, although it is many orders of magnitude faster, since

the estimation of entropic contributions to protein interactions is directly derived from the structure, using a statistical thermodynamics approach.

This makes FoldX a very good tool for these kinds of tasks. Furthermore, this software is constantly developing and new versions keep being released, bearing new options and improved features.

4.1.3 Changing TEV protease Specificity

As shown in this dissertation, the design of proteases "a la carte" is a very exciting challenge and some efforts have been done to focus protein design in this direction. Herein, it was found that the *in silico* process to redesign the cleaving pocket of TEV protease was a major challenge. The general strategy was to divide the problem into small ones, so that each position could be treated individually. Even so, these small jobs turned out to be significant challenges, especially the change of the glutamine from the canonical substrate sequence of the TEV protease, at the cleavage position 307.

4.1.3.1 Scanning the Cleavage Position *in silico*

To reduce the computational time and to make the redesign of the individual positions achievable, some additional assumptions to construct position specific scoring matrices were made: one of the positions affected in the TEV protease was scanned while the other positions remained as Ala. This method works quite well to obtain the best theoretical residues per position and to reject energetically impossible combinations, thus saving a lot of computational time, production efforts and analytical assays. However, this simplification can strongly affect the selection of the putative residues that will be combined later, and they will ultimately have to function simultaneously (see also below). This is one of the critical steps for successful redesigns and is probably the most

responsible for the low activity and loss of specificity of the assayed mutant TEV proteases that is observed here.

4.1.3.2 Reference Selection

Another important step after the *in silico* screening is to compare the energetic data between the reference and mutants. As was mentioned previously, to achieve a successful design the reference template must be the most accurate possible. In this work, the reference templates were chosen for each individual position in the substrate, with affected positions in the TEV enzyme mutated to Ala. This, a priori, seemed to be the most reliable comparative method to evaluate mutants against specific references.

Although the above is true when all affected positions in the TEV protease are simultaneously combined to construct mutants, it should be kept in mind that the scoring matrices for individual positions were built separately, while the other implicated positions were alanines, and that this can preclude measuring the real contribution of an amino acid in the scanning position. In this sense, it is possible that the selected residues per position were suboptimal and that the subsequent construction of the mutants, by combining affected positions, was far from the expected accuracy.

4.1.3.3 Structural strategy validation

As described above, a manual visual checking of the best designs must always be carried out. This step permits the elimination of biologically incorrect structures and allows us to recover rationally those amino acids that were discarded earlier, during the position scanning. For instance, one amino acid can reflect locally suboptimal energies but maybe has more sense in this position in terms of biochemical properties and interactions with the global environment. Thus, a potentially good residue could be disposed of early in the design process and thus some possibilities can be lost. Therefore

this problem can be overcome during visual inspection of the structures, although it requires an experienced user.

In fact this happened during the TEV protease redesign process, with the design 1 (ASRK) and 2 (ATRK). An alanine in position 146, from designs 1 and 4, was included in the combinations chosen for the production and *in vitro* testing of mutants (although Ala was not selected as the optimal amino acid by the computer).

Nevertheless, these selections suffer from being subjective and the evidence can be misinterpreted, thus increasing even more the uncertainties, rather than favoring the design process. In order to try to avoid these sources of inaccuracy, it was asked whether the intramolecular clash energy-value, between enzyme-substrate complexes, can be used as an additional parameter to correct or to normalize the theoretical $\Delta\Delta G$ values of interaction. The reason behind this was that these structures, having Ala in position T146, showed less intramolecular clashes than a Thr in this position (see Tables 3.2 and 3.3). Thus, the addition of values of intramolecular clashes and binding energies of all designs can help to discriminate between possibilities, during the selection of which residues to combine in the final designs.

4.1.3.4 Enzymatic Activity of the Mutant TEV proteases

As discussed above, the possibility of designing customized proteases is a great challenge that can report great benefits. However, some mutants assayed (ASRK, ATRK, MSLK and MTLK) showed neither activity nor cross-reactivity, probably indicating that the enzyme had become inactive after mutagenesis. The mutants MSLK and MTLK correspond to hydrophobic solutions of the P_1 pocket. The purification of these mutants followed the same protocol as that of the wild type TEV protease, and the expression levels were similar. The ratio of soluble fraction versus inclusion bodies was also similar to that of wild-type TEV protease. The absence of activity in these mutants could be explained if the expected complementarities of the residues Met146 and Leu167 in the P_1 pocket (see Fig.3.5 panel D) were not actually possible, producing small conformational

changes that open the binding pocket and preclude substrate binding. Alternatively, if these residues were tightly packed, the conformational change could occur in the opposite direction, closing the substrate binding pocket, and also avoiding binding. In addition to this, it is also possible that the Asp 307 in the substrate can not be compensated properly by Lys174, making the complex unstable.

The mutants ATRK and ASRK correspond to the charged solution of the P_1 pocket. Structurally, the inactive mutant ATRK differs from ASRK only in an extra methyl group of the side chain of Thr148. The theoretical energies of both mutants are similar (see Tables 3.2 and 3.3), as well as the expression levels and other purification conditions. Looking into ATRK in detail, maybe this small clash accounts for destabilizing the substrate-enzyme complex, giving rise to an inactive enzyme. All of these suggested subtle effects of the mutations on the TEV protease give an idea of the difficulty of the redesign exercise.

The partially successful result was the Design 1 TEV protease. ASRK was the mutant that showed low activity to the new substrate Q307D while keeping high cross-reactivity to the wild-type substrate. The reason behind the unsuccessful designs could be the absolute requirement for Gln at position P_1. This requirement has been reported for wild-type TEV protease (Phan et al., 2002; Dougherty et al., 1989a) and for the human rhinovirus 3C protease (Matthews et al., 1999). In addition, the P_1 pocket is located in the close vicinity of the catalytic triad (His46, Asp81 and Cys151), and maybe this fact complicates any effort to alter the specificity at this position.

Although the redesigned P_1 pocket presents a substantial challenge from an engineering standpoint, it was unfortunately not possible to obtain a non-promiscuous mutant TEV protease that was highly specific for the substrate with aspartic acid in this pocket.

4.2 Re-engineering TEV protease: Future Prospects

In order to improve the strategy on redesigning TEV protease specificity, another level of accuracy to consider would be, for instance, to combine more than one key position at the same time. In this case, the possibilities to obtain a very specific binding pocket would be higher, because contextual effects would be considered. Unfortunately, this procedure would be very difficult and would take even longer, because it would need much more computational time and a huge effort for data analysis and the corresponding production and *in vitro* tests.

Even using the old version of FoldX, used in this thesis, this approach would be more precise. However, the protein design software is already more accurate than was previously the case, and it would be worth doing this proposed exercise with a future updated version of FoldX.

Finally, it is important to emphasize that the redesign region might have been too close to the active site. To date, the design of enzyme pockets has not been very successful. With hindsight we should have chosen a position further away.

4.3 Meganucleases: Increasing the Choice

The ability to design obligatory heterodimeric meganucleases could provide a solution for a major specificity and toxicity issue in the genome engineering applications associated with meganucleases. Today, the making of artificial endonucleases with tailored specificities has paved the way for novel approaches in several fields, including gene therapy. For example, meganuclease-induced recombination could be used for the correction of mutations responsible for monogenic inherited diseases (see 1.5.1); the meganuclease can cleave specifically a long sequence in the mutated gene and, by means of the cellular repair mechanism and homology recombination with the wt allele, the gene

can be repaired. This strategy has the advantage to bypass the odds associated with current strategies of random insertion of a complementing transgene (Paques and Duchateau, 2007).

4.4 Meganucleases Versus Zinc Finger Nucleases

As well as meganuclease-directed homology repair, there are other possibilities available, because several reports have shown that engineered Zinc-Finger Nucleases (ZFNs) can trigger efficient site directed recombination in mammalian cultured cells, plants and insects (Durai et al., 2005; Porteus, 2006). However, the low specificity of many of these proteins remains a major issue. Zinc Finger-derived nucleases have proven to be toxic in Drosophila species (Bibikova et al., 2002; Bibikova et al., 2003) and in mammalian NIHT3 mesenchymal cells (Alwin et al., 2005; Porteus and Baltimore, 2003; Porteus and Carroll, 2005), a genotoxic effect that is probably due to frequent off-site cleavage (Porteus, 2006). Although meganucleases have been shown to be less toxic (probably because of better specificity) by different groups (Alwin et al., 2005; Porteus and Baltimore, 2003; Porteus and Carroll, 2005), they can still be harmful at very high doses (Gouble et al., 2006).

The fact of designing obligatory heterodimeric homing endonucleases (HEs) could thus provide an excellent solution for genome engineering applications, because many different engineered monomers could then be combined to target a wide range of DNA sequences, with relatively low toxicity.

4.5 Heterodimer Design Versus Single Chain Solutions

The promiscuity of meganuclease mixtures can be solved by the suppression of homodimer formation. However, this outcome could, in theory, also be achieved by the fusion of the two monomers in a single chain molecule (Chevalier et al., 2002; Epinat et

al., 2003). Unfortunately, this kind of design is relatively dangerous, and can result in badly folded proteins (Epinat et al., 2003).

Herein, by using the protein design algorithm FoldX, the interaction surface of the I-CreI meganuclease was successfully redesigned to obtain a functional obligatory heterodimer. The engineering of such heterodimers is an alternative that provides functional, well folded proteins. Already, hundreds of homodimeric I-CreI derivatives with locally altered specificity have been described in previous reports (Arnould et al., 2006; Smith et al., 2006), and it has been shown that such proteins could be co-expressed to form homo- and heterodimer mixtures. However, the possibility to combine these proteins into obligatory heterodimers will considerably improve the ability to engineer very specific reagents for genome engineering. For therapeutic applications, which require a minimal genotoxicity, this gain in specificity removes one of the last hurdles in the way of using meganucleases for gene therapy and other applications.

4.6 Engineering *in silico*: Final Perspectives

Because the computational method used in this study is based on generally applicable principles, and has been successfully tested on a variety of proteins, this methodology can be further applied to boost the design of other proteins with improved characteristics; specifically, these methods might be easily extrapolated to other proteases and endonucleases.

UNIVERSITAS
Miguel
Hernández

V. CONCLUSIONS

1- A complete methodology on computer-aided protein design has been described and this process could also be used for "a la carte" redesign on other proteins of interest.

2- For first time enzymes have been modeled using FoldX to acquire new unique properties.

3- Mutant TEV proteases, designed *in silico* to recognize a different residue in the cleavage position of the substrate canonical sequence, were not completely successful. The *in vitro* tests of the ASRK TEV protease showed high promiscuity and low activity against the new target site. The close proximity of the P_1 pocket to the active site is probably the cause of the low mutant activity.

4- New meganucleases have been designed *in silico* to create obligatory and specific heterodimeric I-CreI enzyme variants. The *in vitro* tests showed that the aim was achieved, as long as the two different monomers are co-expressed. Hence, a new meganuclease solely recognizing a non-palindromic target site has been obtained for the first time.

5- FoldX has been successfully used as a protein design software, for *in silico* screening of enzyme-substrate and protein-protein interactions.

UNIVERSITAS
Miguel
Hernández

VI. MATERIALS & METHODS

6.1. Computational Protein Design

FoldX vs 2.65 was run in the powerful IBM cluster with 616 CPUs running 64bit Linux from the European Molecular Biology Laboratory [EMBL].

6.1.1 FoldX: a Protein Design Software

The computational protein design algorithm FoldX (http://foldx.embl.de), is an empirical force field that was developed for the rapid evaluation of the effect of mutations on the stability, folding and dynamics of proteins and nucleic acids (Schymkowitz et al., 2005b). This software calculates the free energy of unfolding (ΔG_{total}) of a protein complex or target protein, using the physical description of the atomic interactions with empirical data obtained from experiments on proteins (Guerois et al., 2002; Schymkowitz et al., 2005b). Force field components such as hydrogen bond energies, electrostatics in the complex and their effects on the rate of association (k_{on}), polar and hydrophobic solvation energies, van der Waals' interactions, van der Waals' clashes, and backbone and side chain entropies, are calculated by evaluating the properties of the crystallographic structure: the water accessibility of its atoms and residues, the backbone dihedral angles, the atomic contact map, the hydrogen bond network, and the electrostatic network of the protein (Schymkowitz et al., 2005a). Water molecules making two or more hydrogen bonds with the protein are also taken into account (Schymkowitz et al., 2005c).

6.1.1.1 Side chain Placement Algorithm

The FoldX version 2.65 for Linux used in this work is a fast and accurate energy function that uses a minimum of computational resources. This software performs amino acid (aas) mutations and accommodates the new residue and its surrounding aas in the following way: It first mutates the selected position to alanine and annotates the side chain energies of the neighboring residues. Then, it mutates the alanine to the selected aa

and recalculates the side chain energies of the same neighboring residues. Those residues that exhibit an energy difference are mutated to themselves to see whether another rotamer it is more favorable in this position. This option ensures that whenever FoldX is mutating a protein, or DNA, it always moves the same neighbors in the wild type (wt) and in the mutant, producing for each mutant a corresponding PDB for the wt.

6.1.1.2 Force Field Description

The FoldX force field calculates the free energy (ΔG in kcal mol^{-1}) of unfolding of a target protein, following the equation of empirical terms that have been found to be important for protein stability:

$$\Delta G = a \cdot \Delta G_{vdw} + b \cdot \Delta G_{solvH} + c \cdot \Delta G_{solvP} + d \cdot \Delta G_{wb} + e \cdot \Delta G_{hbond} + \\ + f \cdot \Delta G_{el} + g \cdot \Delta G_{Kon} + h \cdot T\Delta S_{mc} + k \cdot T\Delta S_{sc} + l \cdot \Delta G_{Clash}$$

In this expression, (*a*...*l*) are relative weights of the different energy terms used for the free energy calculation. ΔG_{vdw} is the sum of the van der Waals contributions of all atoms with respect to the same interactions with the solvent. ΔG_{solvH} and ΔG_{solvP} is the difference in solvation energy for apolar and polar groups respectively when going from the unfolded to the folded state. ΔG_{wb}, is the extra stabilizing free energy provided by a water molecule making more than one hydrogen-bond to the protein (water bridges) that cannot be taken into account with non-explicit solvent approximations (Petukhov et al., 1999). ΔG_{hbond} is the free energy difference between the formations of an intra-molecular hydrogen-bond compared to inter-molecular hydrogen-bond formation (with solvent). ΔG_{el} is the electrostatic contribution of charged groups, including the helix dipole. ΔS_{mc} is the entropy cost for fixing the backbone in the folded state; this term is dependent on the intrinsic tendency of a particular amino acid to adopt certain dihedral angles (Munoz and Serrano, 1994b). Finally ΔS_{sc} is the entropic cost of fixing a side chain in a particular conformation (Abagyan and Totrov, 1994; Munoz and Serrano, 1994b).

Also the interaction with the solvent is treated in two steps: first, the bulk solvent is treated as a desolvation term that is continuously scaled with the burial of an atom and separated into contributions from hydrophobic (ΔG_{solvH}) and polar (ΔG_{solvP}) groups. These solvation parameters have been derived from experiments in which amino acids are transferred from water to an organic solvent; this is assumed to mimic the transition that is experienced by an aa during folding, from solvent exposure in the unfolded state to burial in a hydrophobic environment in the native state. In addition, those water molecules that have a persistent interaction with groups of the protein, (waters that make more than two hydrogen bonds with the protein), are calculated explicitly in the ΔG_{wb} term (Abagyan and Totrov, 1994; Guerois et al., 2002; Munoz and Serrano, 1994b; Petukhov et al., 1999). The combination of a continuous solvation scale with an explicit consideration of the essential water molecules allows fast calculations while providing essential details. The van der Waals' terms, ΔG_{vdw}, are calculated in a similar fashion to the desolvation, but taking into account experimental transfer energies from water to vapor. Hydrogen bonds are calculated on the basis of simple geometric considerations and their energy, ΔG_{hbond}, is inferred from protein engineering double mutant cycles. The electrostatic contribution to the free energy, ΔG_{el}, is calculated from a simple implementation of Coulomb's law, in which the dielectric constant is scaled with the burial of the bond under consideration. Hypothetical atoms are included in the calculations of the Coulombic interactions in order to capture some specific aspects of protein stability:

(i) charged atoms are placed at the N- and C-terminal of each α-helix, to obtain some measure of the helix dipole interaction.

(ii) aromatic rings carry positive charges on the edges and negative charges above the centre of the ring.

In the case of protein complexes, an additional electrostatic contribution is calculated between the atoms of different polypeptide chains, ΔG_{kon}, based on the empirical equation of Schreiber *et al*., which was shown to give a good estimation of the

association rate (k_{on}) of complex formation (Selzer et al., 2000). An important difference between FoldX and other force fields is the crude entropy estimation that is used to obtain a measure of the free energy. Entropy calculations usually involve large simulations of the conformational freedom of the side chains and the backbone of the protein. In FoldX the entropic penalty for fixing the backbone in a given conformation, ΔS_{mc}, is derived from a statistical analysis of the phi–psi distribution of a given Aa, as observed in a set of non-redundant high-resolution crystal structures. This entropy is scaled by:

(i) the accessibility of the main chain atoms.
(ii) the energy of hydrogen bond interactions made by the corresponding residue or its direct neighbors.

The entropy cost of fixing a side chain in a particular conformation, ΔS_{sc}, is obtained by scaling a set entropy parameters calculated by Abagayan and co-workers (Abagyan and Totrov, 1994) to the burial of the side chain. Finally, the ΔG_{clash} term provides a measure of the steric overlaps between atoms in the structure. There are two methods of incorporating the resulting repulsive energy into the FoldX calculation:

(i) when analyzing point mutations it is recommended to use a soft penalization of the van der Waals' clashes
(ii) when doing protein design, full penalizations should be applied.

The highest accuracy in the FoldX predictions is achieved when the energy difference can be calculated between two well-defined structures, such as between the wt and a mutant, or between the bound and unbound forms of a protein complex (to determine the binding free energy). The difference in the calculated free energies ($\Delta\Delta G$) between the final state (the mutant) and the reference state (the wt protein) correlates well with the experimentally observed change in stability.

Indeed, FoldX was calibrated using experimental mutational free energy changes from a collection of more than 1000 point mutants, covering many different proteins

(Guerois et al., 2002), and in its current release yields a correlation of 0.81 with a SD of 0.46 kcal mol^{-1} between calculated and experimental $\Delta\Delta G$s.

On the other hand, the free energy of folding is calculated from the difference in Gibbs free energy between the detailed three-dimensional structure found in the PDB file and a hypothetical unfolded reference state of which no structural detail is available. The main assumption in this approach is the absence of persistent structure in the denatured state, which in a range of proteins was experimentally shown to be only partly correct.

Therefore, although the free energy for folding predicted by the FoldX force field for most small protein domains yields a number between –5.0 and –15 kcal/mol^{-1}, this value should not be considered as absolute since it could have large error. Nevertheless, positive energies are normally indicative of problems with the structure under scrutiny and as such, one should bear in mind, when using FoldX, that the best results are obtained when comparing known structures.

6.1.1.3 Description of FoldX Commands and Options

<RepairPDB>

Identifies those residues which have bad torsion angles, or van der Waals' clashes, or total energy and then FoldX 'mutates' to themselves, to improve these problem areas. The way it operates is the following: First FoldX mutates the selected position to Ala and annotates the side chain energies of the neighbor residues. Then it mutates the Ala back to the original aa and re-calculates the side chain energies of the same neighbor residues. Those that exhibit an energy difference are then mutated to themselves to see if another rotamer will be more favorable.

<BuildModel>

This command mutates protein or DNA, and simultaneously moves the same neighbors in the wt and in the mutant, thus producing for each mutant a corresponding PDB for its wt template. This is because each mutation will move different neighbors and therefore needs different wt references.

<AnalyseComplex>

This determines the interaction energy between 2 molecules or a group of molecules. The way it operates is by unfolding the selected targets and determining the stability of the remaining molecules and then subtracting the sum of the individual energies from the global energy.

$$Eb = Es - \sum (EsA + EsB)$$

The output file contains the different energy terms (all of them reflecting changes in the respective energies upon binding) plus an extra one showing intrachain clashes of the residues involved in the interface. This term is important when designing protein complex interfaces since it could help to correct solutions where a residue has a very good interaction with the neighbor chain, but is in a very strained conformation with respect to its own chain. Thus that conformation is not realistic. The users can select which side chain or group of side chains they want to use to determine the interaction energy with the rest of the protein. There are two options that can be used with this command:

<complex_with_DNA>; when activated, this automatically divides the PDB into two groups, DNA and Protein, and calculates the interaction between both.

<optimize_complex_domains>; when activated, the selected molecule in isolation is optimized, as well as the complex.

6.2 In silico Studies

6.2.1 TEV protease Designs: Key Residue Scanning Strategy

The X-ray structure of the catalytically inactive TEV protease (mutation C151A) complexed with its canonical decapeptide target sequence (T**E**NL**Y**F**Q**//SGT), determined at 2.20Å resolution (PDB: 1lvb)(Phan et al., 2002), was used as template to redesign the TEV protease binding site. The underlined amino acids are the key residues that the protease recognizes to locate the cleavage point (slashes) in the active center for catalysis.

In order to redesign the recognition specificity of the TEV protease, the following scanning strategy was adopted:

The original 3D structural crystallographic coordinates file was first edited and cleaned using the SwissPdb Viewer software (Guex, N. and Peitsch, M.C., 1997, *Electrophoresis,* 18, 2714-2723; http://www.expasy.org/spdbv/). Only one TEV protease chain and one substrate sequence chain were left. The resulting simplified structure was optimized using the <RepairPDB> command of FoldX, in order to release any van der Waals clashes, and this was used as the reference template for the designs (from now on called "TEV_repaired").

The global strategy to choose the target position on the substrate, to execute the TEV protease redesign, is shown in the results section (see 3.1.1.1). Below, the computational steps are described.

6.2.1.1 Substrate Global Scanning Position

First, the selected key recognition positions (E302, Y305 and Q307) were treated individually to simplify the design.

Second, all 20 natural amino acids were constructed in every key position, giving 20 structures per position. The resulting structures were analyzed using the <AnalyseComplex> command.

Third, for each mutated structure in the substrate key position, the residues interacting with the protease were chosen by rational design, using the SwissPdb Viewer software for looking at the interface, and selecting those positions that were directly involved in binding.

This is a global scanning strategy that aims to cover all possible key positions substrate sequences and the corresponding positions to be redesigned on TEV protease, for each specific ones. As a result, the chosen key position to mutate was Q307.

6.2.1.2 Redesigning TEV protease to Cleave Q307D substrate (Asp on P_1; the Cleavage Site)

For the Q307D substrate, the Asp interacts with T146, D148, H167 and N174 residues in the protease. These positions of interest were then mutated to Alanine (T146A, D148A, H167A and N174A) using <RepairPositionScan>, and all 20 aas were constructed on these positions, one by one (A146x, A148A, A167A and A174A), using the <BuildModel> command. Thus, this step generated a matrix of 80 structures (20 aa x 4 positions) that were analyzed by the <AnalyseComplex> command to compute the different interaction energies. These data were tabulated and sorted by increasing energy so that the lower the energy difference with the wt template, the better the interaction was (see 3.1.1.4). The tables, together with the visual inspection of the resulting structures, allowed choosing the best amino acid(s) to combine in the TEV designs for this key position (for Q307D were T146(MSTV), D148(ST), H167(RMLFI) and N174(MKLHQ); see Table 3.1. The selection of a small set of aa per position allows the mutagenesis of all positions simultaneously. The combination of the mentioned aas per position, using <BuildModel> and <AnalyseComplex> commands, renders the energy and structure of 200 designed TEV proteases for this Q307D key position mutation. Again, these resulting

energy tables, and the inspection of the structures, help to discriminate the best designs to be produced and purified on the bench.

6.2.2 I-CreI Heterodimers design

The different heterodimers were designed and evaluated also using FoldX (see 6.1.1). The X-ray structure of the I-*Cre*I homodimer determined at 2.05 Å resolution (PDB: 1g9y), bound to its cognate DNA target sequence (Chevalier et al., 2001a; Phan et al., 2002), was used as a template to design the heterodimeric interface of I-*Cre*I.

The structure was first optimized using the <RepairPDB> command of FoldX, in order to release van der Waals clashes, and each position of interest (chosen rationally using SwissPDB Viewer) was mutated to Alanine (<BuildModel> command). All models (heterodimers and homodimers alike) were generated separately, and each model of the complex was analyzed through the <AnalyseComplex> command to compute the different interaction energies.

6.3 DNA Cloning and Site-Directed Mutagenesis

All oligos and reagents were of analytical grade from Sigma and all DNA modifying enzymes were from New England Biolabs.

6.3.1. TEV Protease and substrate mutants

6.3.1.1 TEV Protease mutants

A pET23(+) vector carrying the wt TEV protease obtained from Dr. Ario Di Marco (EMBL-Heidelberg, Protein Expression Facility) was used as template for mutations that were introduced following the protocol of the QuikChange® Site-Directed Mutagenesis Kit (Stratagene) with the following oligos pairs in each case:

- **Design 1**; (A146, S148, R167, K174) TEV protease

(i) TEV_1_AS_frw; C TGG AAG CAT TGG ATT CAA GCG AAG AGC GGG CAG TGT GGC AGT CC. and reverse-complement oligo.

(ii) TEV_RK_frw; G TTC ATT GTT GGT ATA CGT TCA GCA TCG AAT TTC ACC AAA ACA AAC AAT TAT TTC. and reverse-complement oligo.

- **Design 2**; (M146, S148, L167, K174) TEV protease

(iii) TEV_2_MS_frw; C TGG AAG CAT TGG ATT CAA ATG AAG AGC GGG CAG TGT GGC AGT CC. and reverse-complement oligo.

(iv) TEV_LK_frw; G TTC ATT GTT GGT ATA CTG TCA GCA TCG AAT TTC ACC AAA ACA AAC AAT TAT TTC and reverse-complement oligo.

- **Design 3**; (M146, T148, L167, K174) TEV protease

(v) TEV_3_MS_frw; C TGG AAG CAT TGG ATT CAA ATG AAG ACC GGG CAG TGT GGC AGT CC, reverse-complement oligo and (iv) oligos

- **Design 4**; (A146, T148, R167, K174) TEV protease

(vi) TEV_4_AT_frw; C TGG AAG CAT TGG ATT CAA GCG AAG ACC GGG CAG TGT GGC AGT CC, reverse-complement oligo and (ii) oligos

These PCR reactions were used to transform *Escherichia coli* XL1-Blue super competent cells according to the manufacturer's instructions (Stratagene). The fidelity of five clones of each construct was confirmed by sequencing using the primer for T7-terminal-rev (EMBL-Heidelberg, Gene Core Facility).

6.3.1.2 Substrate-Reporter Constructions

The pGFPmut3.1 vector (Clontech) was used as a template to amplify, by polymerase chain reaction (PCR), the coding sequence of the reporter, green fluorescent mutant 3 protein (GFPmut3), using primers to give a product flanked by NcoI and BamHI restriction enzyme (REs) sites. This PCR reaction product was run on a 1% agarose gel and the correct band was excised and purified by QIAquick® gel extraction kit (Quiagen). The pETM30 vector (EMBL-Heidelberg, Protein Expression Facility) coding for GST in fusion with TEV-site was cleaved with the above mentioned REs in order to exchange the Dimerization Cofactor of HNF-1(DCoH) with GFP. The PCR fragment was digested in the same way, ligated during 1h at room temperature (3 fold excess of insert over vector, T4 ligase) to create the fusion: GST-Substrate_TEV_Site-GFP (see Fig.6.1). This reaction was used to transform *E.coli* XL1-Blue supercompetent cells and the pellets were processed to purify the mutated plasmids. The fidelity of five clones of each construct was confirmed by sequencing using the primer pSubsTEV_rev: CTG CGG ATC CTA TTT GTA CAG TTC ATC CAT GCC ATG TGT AAT CC.

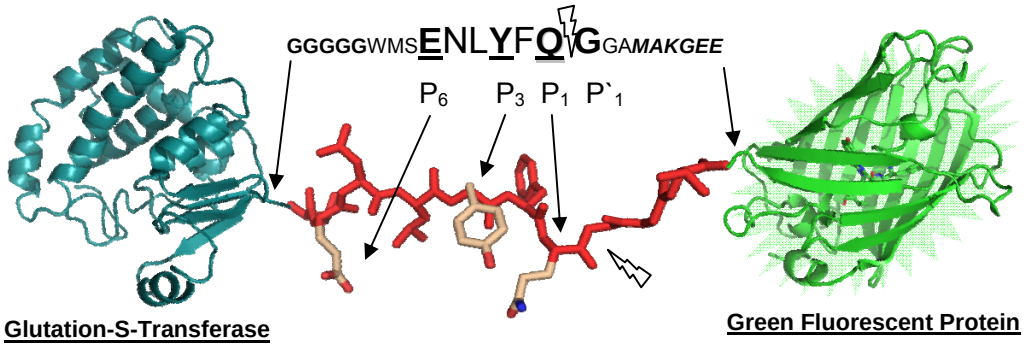


Figure 6.1 The substrate-reporter. *TEV protease substrate-reporter used in the cleavage assays with the canonical sequence between GST and GFP.*

6.3.2 I-CreI Meganuclease Mutants

6.3.2.1 Generation of the KTG and QAN Meganucleases

The scaffold meganucleases used to design the specific obligate heterodimers in this study, were two previously engineered homodimeric variants of I-CreI meganuclease monomers, that recognize different DNA sequences (provided by Cellectics S.A).

As described previously (Arnould et al., 2006; Chames et al., 2005; Epinat et al., 2003; Smith et al., 2006), these proteins were engineered using a High Throughput Screening of an D75N I-CreI library, having random variations at positions Q44, R68 and R70, with 64 palindromic DNA targets, resulting from base substitutions in positions ±3, ±4 and ±5 of a 22 bp palindromic target, cleaved by I-*Cre*I (see 6.3.2.3).

One I-CreI variant, having mutations Q44K, R68T and R70G (denoted as "KTG") recognizes the bases CCT at positions -5, -4 and -3 of the DNA target. The other I-CreI meganuclease variant, having mutations Q44Q, R68A and R70N (called "QAN") recognizes the bases GTT at positions -5, -4 and -3 of the DNA target. The experimental cleavage data of these enzymes were also validated with homology models of the complexes using FoldX; as described previously (Arnould et al., 2006).

6.3.2.2. Cloning Meganuclease Mutants

The two homodimerising meganucleases KTG and QAN, based on the I-*Cre*I meganuclease scaffold, were each mutated at up to 6 amino acid positions to form two compatible heterodimerising interfaces, denoted KTG-A2 and QAN-B3. Mutations were introduced as described by Quikchange® kit protocol.

KTG-A1 and QAN-A1 mutations (K7R, E8R, K96R and L97F) were introduced using these three complementary primer sets:

(i) A1_FLR_frw; CTG GAC AAA CTA GTG GAT AGA ATT GGC GTT GGT TAC G.

(ii) A1/A2_RR_frw; CAA TAC CAA ATA TAA CAG GCG GTT CCT GCT GTA CCT GGC CG.

(iii) A1/A2_RF_frw; TCA ACT GCA GCC GTT TCT GAG ATT CAA ACA GAA ACA GGC AAA CC.

KTG-A2 and QAN-A2 mutations (K7R, E8R, F54W, E61R, K96R and L97F) were introduced using the complementary primer sets (ii), (iii) and also:

(iv) A2_WLR_frw; CCA GCG CCG TTG GTG GCT GGA CAA ACT AGT GGA TAG AAT TGG CGT TGG TTA CG.

QAN-B3 mutations (K7E, F54G, L58M and K96E) were introduced also using three complementary primer sets:

(v) B3/B4_EE_frw; CAA TAC CAA ATA TAA CGA AGA GTT CCT GCT GTA CCT GGC CG.

(vi) B3/B4_GME_frw; CCA GCG CCG TTG GGG TCT GGA CAA AAT GGT GGA TGA AAT TGG CGT TGG TTA CG.

(vii) B3_EL_frw; TCA ACT GCA GCC GTT TCT GGA ACT GAA ACA GAA ACA GGC AAA CC.

QAN-B4 mutations (K7E, F54G, L58M, K96E and L97G) were introduced also using the complementary primer sets (v), (vi) and also:

(viii) B4_EG_frw**;** TCA ACT GCA GCC GTT TCT GGA AGG GAA ACA GAA ACA GGC AAA CC

For example, the first primer set was used for PCR and, after purifying the PCR product and digesting away parental plasmid DNA with DpnI, for transformation of *E. coli* TOP10 supercompetent cells (Invitrogen). Then approximately 300 transformant bacterial colonies were pooled in 2 ml Luria Broth medium (LB), and plasmid DNA was recovered by miniprep. This DNA was used as template for a second and then a third round of PCR with corresponding mutagenic primers. Five third-round mutants were verified by DNA sequencing using the primer pETSeq_frw: TTG TGA GCG GAT AAC AAT TCC. The same method was used to make the alternative designs for the heterodimer pairs. It is worth noting that the primers above are universal for any I-*Cre*I mutant with altered specificity, since the dimer interface mutations are outside the DNA recognition region.

6.3.2.3. Preparing DNA Target Sites

The sequence of the palindromic target (CCT and GTT), digested by KTG and QAN meganucleases, respectively, are defined from **TCA** to **TGA** (underlined below), with the mutated target nucleotides in positions -5, -4, -3 and 3, 4, 5 (blue italic):

- **Meganuclease Homodimers KTG,** cleave CCT -

-12 -11-10-9-8-7-6 -5 -4 -3 -2 -1 / 1 2 3 4 5 6 7 8 9 10 11 12
TGGCATACAAGTT**TCA**AAAC*CCT*GT/AC*AGG*GTTT**TGA**CAATCGTCTGTCA

- **Meganuclease Homodimers QAN,** cleave GTT -

-12 -11-10-9-8-7-6 -5 -4 -3 -2 -1 / 1 2 3 4 5 6 7 8 9 10 11 12
TGGCATACAAGTT**TCA**AAAC*GTT*GT/AC*AAC*GTTT**TGA**CAATCGTCTGTCA

- **Meganuclease Obligate Heterodimers KTG--QAN,** cleave pseudo-palindromic targets)

-12 -11-10-9-8-7-6 -5 -4 -3 -2 -1 / 1 2 3 4 5 6 7 8 9 10 11 12
TGGCATACAAGTT**TCA**AAAC*GTT*GT/AC*AGG*GTTT**TGA**CAATCGTCTGTCA
or
TGGCATACAAGTT**TCA**AAAC*CCT*GT/AC*AAC*GTTT**TGA**CAATCGTCTGTCA

All targets are in pGEM-T vector (Promega). The targets were cloned according to the manufacturer's instructions. Both are functional and give same pattern for I-*Cre*I digestion profile.

XmnI was used to prepare a linearized solution of each plasmid at 1µg/µl in same buffer than I-*Cre*I (see below). The cleavage and correct linearization was verified by running a 1% Agarose gel with 130 ng of each target site, for 45 min at 110 volts.

6.4. Protein Expression and Purification

6.4.1 Production and Purification of TEV Proteases and Substrates

6.4.1.1 TEV Proteases

A single colony transformant of *E. coli* BL21 (DE3) pLysS (Stratagene), carrying the plasmid coding for the desired TEV protease, was used to inoculate 6 tubes with 4 ml of LB (medium plus 32µg/ml chloramphenicol and 50µg/ml ampicillin). These were grown overnight at 37°C on a shaker; ~220 rpm. This pre-culture was used to inoculate 2 x 0.5 L LB plus appropriate antibiotics (in 2-L flasks). At an OD_{600} of 0.6–0.8, flasks were put on ice for 15 min to arrest growth. Expression was induced by adding 0.1 mM final isopropyl-1-thio-β-D-galactopyranoside (IPTG), shaking at 200 rpm for 16 hours at 20°C. Cells were subsequently harvested by centrifugation (15 min, 16,000 g).

The procedure was carried out at 4ºC thereafter. Pellets were resuspended in 30 ml ice-cold lysis buffer (50mM Tris·HCl pH 8, 200mM NaCl, 5mM $MgCl_2$, 10% Glycerol, 10 mM imidazole) containing 1 unit/µl DNAse I and a tablet of complete protease inhibitor cocktail EDTA-free (Roche). The suspension was immediately frozen in liquid nitrogen and thawed for 16 hours at 4ºC on a rotating platform (60 rpm). The suspension was homogenized with an Ultra Turrax T25 (Jankel & Kunkel, IKA-Labortechnik); 3 cycles of 1 min on ice) and then broken with an EmulsiFlex-C5 homogenizer (Avestin), for 5 rounds of 500-1000 psi (pounds per square inch) each. The lysate was centrifuged at 150,000 g for 60 min. This supernatant was cleared through a 0.45µm filter (Millipore). Using a ÄKTAfplc (GE Healthcare Life Science), a 5ml Hi-Trap column (Amersham-Pharmacia) was loaded with 2 bead volumes (vol) of 250 mM

$NiS0_4$, and rinsed with 3 volumes of binding buffer (50mM Tris·HCl pH 8, 300mM NaCl, 1mM DTT, 20% glycerol, 10mM imidazole). The supernatant was then applied to the column and washed with washing buffer (binding buffer with 50 mM imidazole) until the A_{280nm} returned to its basal level. Protein was eluted with elution buffer (0.3M imidazole). The protein peak was collected and immediately applied to a dialysis membrane (Molecular Weight Cut Off = 12 kDa, Spectra), placed in 2 liters of dialysis buffer (50 mM Tris·HCl pH 8, 200 mM NaCl, 1 mM DTT, 1mM EDTA, 50% glycerol) at 4ºC, for at least 12 hours. The purified protein was aliquoted and snap-frozen in liquid nitrogen and stored at -80ºC.

6.4.1.2 Substrate-reporter Constructions

For expression and purification of the substrates, the same process as with TEV proteases was followed, except for some modifications: all the steps were made at room temperature, and after clearing through a 0.45μm filter (Millipore), the supernatant was passed through to a 5ml GSTrap FF column (Amersham-Pharmacia) equilibrated with 5 vol of binding buffer (PBS pH 7.3; 140 mM NaCl, 2.7 mM KCl, 10mM NA_2HPO_4, 1.8 mM KH_2PO_4, pH 7.3). The column was washed with 10 vol of binding buffer until the A_{280nm} returned to its basal level. Protein was eluted with elution buffer (50mM Tris-HCl, 10mM reduced glutathione, pH 8.0). The affinity-purified protein was aliquoted, snap-frozen in liquid nitrogen and stored at -80ºC.

6.4.2 Expression and Purification of I-CreI Endonucleases

6.4.2.1 Production of Monomers of I-CreI Endonucleases

Fresh BL21(DE3) (Stratagene) transformants carrying the pET plasmid (Novagen) coding for the I-*Cre*I monomers, were grown overnight in 5 ml of LB medium plus 30 μg/ml kanamycin at 37°C on a shaker. This pre-culture was expanded to a larger culture (1:200). At an OD_{600} of 0.6–0.8, flasks were put on ice for 15 min to arrest growth. Expression was induced by adding IPTG (0.1 mM final) for 18 hours at 16°C,

and cells were harvested by centrifugation (15 min, 16,000 g). Pellets were resuspended in 30 ml ice-cold lysis buffer (50mM Tris·HCl pH 8, 200mM NaCl, 5mM $MgCl_2$, 10% Glycerol, 10 mM imidazole) containing 1 unit/µl DNAse I and the procedure was carried out at 4ºC thereafter. The suspension was immediately frozen in liquid nitrogen and thawed for 16 hours at 4ºC on a rotating platform (60 rpm). The suspension was homogenized with an Ultra Turrax T25 (Jankel & Kunkel, IKA-Labortechnik); 3 cycles of 1 min on ice) and then broken with an EmulsiFlex-C5 homogenizer (Avestin), for 5 rounds of 500-1000 psi (pounds per square inch) each. The lysate was centrifuged at 150,000 g for 60 min. This supernatant was cleared through a 0.45µm filter (Millipore). Using a ÄKTAfplc (GE Healthcare Life Science) a 5ml Hi-Trap column (Amersham-Pharmacia), was loaded with 2 bead volumes (vol) of 250 mM $NiSO_4$, and rinsed with 3 volumes of binding buffer (50mM Tris·HCl pH 8, 300mM NaCl, 1mM DTT, 20% glycerol, 10mM imidazole). The supernatant was then applied to the column and washed with washing buffer (binding buffer with 50 mM imidazole) until the A_{280nm} returned to its basal level. Protein was eluted with elution buffer (0.3M imidazole). The protein peak was collected and immediately applied to a dialysis membrane (Molecular Weight Cut Off =3.5 kDa, Spectra), placed in 2 liters of dialysis buffer (50 mM Tris·HCl pH 8, 200 mM NaCl, 1 mM DTT, 1mM EDTA, 50% glycerol) at 4ºC, for at least 12 hours. The purified protein was aliquoted, snap-frozen and stored at -80ºC.

6.4.2.2 Co-expression of the Obligate Heterodimer KTG-A2—QAN-B3

In order to remove the His tag from the QAN-B3 monomer, it was excised from parent plasmid pCLS1214 (pET-series) with NcoI and NotI. This fragment was then cloned into similarly-cut pCDFDuet1 plasmid (Novagen). TOP10 ultracompetent cells (Invitrogen) were transformed with this mixture and selected in 50 µg/ml Streptomycin-Spectinomycin sulphate. Bacterial clones were verified by DNA sequencing.

BL21 (DE3) ultracompetent cells (Stratagene) were co-transformed with 10 ng of each plasmid (pCLS1211-KTG-A2 and pCDFDuet1-QAN-B3). The double transformants were selected by growing the transformed colonies in presence of Kanamycin and Streptomycin-Spectinomycin sulphate. The purification was performed essentially as described above.

6.4.3 Analytical Centrifugation of Meganucleases

The oligomeric state of meganucleases and mutants was investigated by monitoring sedimentation properties in centrifugation experiments; 1.04 mg of pure protein was used per sample (0.52 mg/ml of each monomer or 1.04 mg/ml of individual wt homodimers) in storage buffer (50 mM Tris-HCl pH 8.0, 225 mM NaCl, 1 mm EDTA, 1 mM DTT, 8% glycerol).

The sedimentation velocity profiles were collected by monitoring the absorbance signal at 280 nm as the samples were centrifuged in a Beckman Optima XL-A centrifuge fitted with a four-hole AN-60 rotor and double-sector aluminium centerpieces (48 000 rpm, 4 °C). Molecular weight distributions were determined by the C(s) method (Schuck, P., Biophys., 2000, 78, 1606-1619) implemented in the Sedfit and UltraScan 7.1 software packages [Demeler.B.,2005, http://www.ultrascan.uthscsa.edu].

Buffer density and viscosity corrections were made according to data published by Laue *et al.* (In *Analytical Ultracentrifugation in Biochemistry and Polymer Science, 1992, Harding S.E., Rowe A,J., Horton J.C. Eds, pp. 90-125, Royal Society of Chemistry, Cambridge*) as implemented in UltraScan 7.1.

The partial specific volume of meganucleases and mutants was estimated from the protein sequence according to the method by Cohn E. J. and Edsall J.T. (In *Proteins, Amino Acids and Peptides, 1943, p.157, Reinhold, New York*).

6.5. *In vitro* Assays

6.5.1 General Methods

6.5.1.1 Protein Concentration

Protein concentration was determined by "Bradford" (BioRad), using bovine serum albumin (BSA) as standard. Samples were incubated in 1x Bradford reagent for 5 min at room temperature and OD at 595 nm was measured on a spectrophotometer.

6.5.1.2 Protein Visualization

All the samples from the different purifications and reactions assays were analyzed by sodium dodecylsulfate-polyacrylamide gel electrophoresis (SDS-PAGE) technique (Maizel, 2000), using 15% Tris-HCl "Criterion Precast System" gels (BioRad). The samples were denatured at 90ºC for 15 min and then loaded on the gels and run for 60 min at 160 volts. Proteins were visualized by staining with Coomassie Brilliant Blue (Biorad) during 30-45 min and then destained with 10% glacial acetic acid, until all the background staining was removed. Gels were scanned at high resolution.

6.5.2 *In vitro* Cleavage Assays of TEV Proteases

The purified TEV proteases and substrate-reporters were diluted to 1µg/µl in fresh dialysis or elution buffer respectively (see above). In each case, the experimentally determined optimal cleavage conditions for the enzymes were used: 1:50 enzyme/substrate ratio was mix with reaction buffer (50 mM Tris-HCl ph=8.0, 150mM NaCl, 0.5 mM EDTA and 1 mM DTT) at room temperature (22-25ºC).

6.5.3 Quantification of TEV protease Activities

The SDS-PAGE gels from the kinetic assays were scanned analyzed using ImageJ, a public domain Java image processing program (http://rsb.info.nih.gov/ij/download.html).

The method followed to analyze one-dimensional electrophoretic gels is described in (http://rsb.info.nih.gov/ij/docs/index.html). The resulting data were used to plot the substrate cleavage as a function of time. The amount of enzyme was taking into account for these calculations.

6.5.4 DNA Digestion Assays for I-CreI Meganucleases

Cleavage of the target sequences was determined as previously described (Epinat *et al*., Nucleic Acids Res., 2003, 31, 2952-2962) with modifications: co-expressed, purified enzymes were diluted to 1 µg/µl in fresh dialysis buffer (in the case of the designed monomers which were purified separately, 1.5 µg of each monomer was added, they were brought to 0.5 µg/µl each). Enzymes were stored at –80˚C. The reaction mixture was prepared using 3.75 µM enzyme, 34 nM of purified 3.2 kb DNA plasmid containing the appropriate target sequences (pre-linearized with *Xmn*I) in digestion buffer consisting of: 25 mM HEPES (pH 8), 5 % Glycerol, 10 mM $MgCl_2$, and 50 mM NaCl or NaCl concentrations range between 50 (low ionic strength) and 300 mM NaCl (high ionic strength); in the case of co-expressed KTG-A2—QAN-B3: 225 mM NaCl, in a 20 µl final reaction volume. The digestion mixtures were incubated for 60 min at 37ºC in a water bath and then mixed with 2.5 µl volume of Stop buffer, modified from Wang *et al.*, Nucleic Acids Res., 1997, 25, 3767-3776 (50% Glycerol, 0.1 M EDTA, 0.5 % SDS, 1mg/ml Proteinase K, 0.25 % bromophenol blue). Samples were incubated for 30 min more at 37˚C, and then half of each sample was visualized on a 1 % agarose gel.

UNIVERSITAS
Miguel
Hernández

VII. BIBLIOGRAPHY

Abagyan,R. and Totrov,M. (1994). Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. J. Mol. Biol. *235*, 983-1002.

Alwin,S., Gere,M.B., Guhl,E., Effertz,K., Barbas,C.F.3., Segal,D.J., Weitzman,M.D., and Cathomen,T. (2005). Custom zinc-finger nucleases for use in human cells. Mol Ther *12*, 610-7.

Anfinsen,C.B. (1973). Principles that govern the folding of protein chains. Science *181*, 223-230.

Arnould,S., Chames,P., Perez,C., Lacroix,E., Duclert,A., Epinat,J.C., Stricher,F., Petit,A.S., Patin,A., Guillier,S., Rolland,S., Prieto,J., Blanco,F.J., Bravo,J., Montoya,G., Serrano,L., Duchateau,P., and Paques,F. (2006). Engineering of large numbers of highly specific homing endonucleases that induce recombination on novel DNA targets. J. Mol. Biol. *355*, 443-458.

Ashworth,J., Havranek,J.J., Duarte,C.M., Sussman,D., Monnat,R.J., Jr., Stoddard,B.L., and Baker,D. (2006). Computational redesign of endonuclease DNA binding and cleavage specificity. Nature *441*, 656-659.

Baker,D. and DeGrado,W.F. (1999). Engineering and design. Curr. Opin. Struct. Biol. *9*, 485-486.

Baker,N.A. (2005). Improving implicit solvent simulations: a Poisson-centric view. Curr. Opin. Struct. Biol. *15*, 137-143.

Balakin,K.V., Kozintsev,A.V., Kiselyov,A.S., and Savchuk,N.P. (2006). Rational design approaches to chemical libraries for hit identification. Curr. Drug Discov. Technol. *3*, 49-65.

Balbin,M., Fueyo,A., Tester,A.M., Pendas,A.M., Pitiot,A.S., Astudillo,A., Overall,C.M., Shapiro,S.D., and Lopez-Otin,C. (2003). Loss of collagenase-2 confers increased skin tumor susceptibility to male mice. Nat. Genet. *35*, 252-257.

Berg,D.T., Gerlitz,B., Shang,J., Smith,T., Santa,P., Richardson,M.A., Kurz,K.D., Grinnell,B.W., Mace,K., and Jones,B.E. (2003). Engineering the proteolytic specificity of activated protein C improves its pharmacological properties. Proc. Natl. Acad. Sci. U. S. A *100*, 4423-4428.

Berman,H., Henrick,K., Nakamura,H., and Markley,J.L. (2007). The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Res. *35*, D301-D303.

Bibikova,M., Beumer,K., Trautman,J.K., and Carroll,D. (2003). Enhancing gene targeting with designed zinc finger nucleases. Science *300*, 764.

Bibikova,M., Golic,M., Golic,K.G., and Carroll,D. (2002). Targeted chromosomal cleavage and mutagenesis in Drosophila using zinc-finger nucleases. Genetics *161*, 1169-75.

Blanes-Mira,C., Ibanez,C., Fernandez-Ballester,G., Planells-Cases,R., Perez-Paya,E., and Ferrer-Montiel,A. (2001). Thermal stabilization of the catalytic domain of botulinum neurotoxin E by phosphorylation of a single tyrosine residue. Biochemistry *40*, 2234-2242.

Blobel,C.P. (2000). Functional processing of fertilin: evidence for a critical role of proteolysis in sperm maturation and activation. Rev. Reprod. *5*, 75-83.

Bolduc,J.M., Spiegel,P.C., Chatterjee,P., Brady,K.L., Downing,M.E., Caprara,M.G., Waring,R.B., and Stoddard,B.L. (2003). Structural and biochemical analyses of DNA and RNA binding by a bifunctional homing endonuclease and group I intron splicing factor. Genes Dev *17*, 2875-88.

Bolon,D.N., Marcus,J.S., Ross,S.A., and Mayo,S.L. (2003). Prudent modeling of core polar residues in computational protein design. J. Mol. Biol. *329*, 611-622.

Bolon,D.N. and Mayo,S.L. (2001). Enzyme-like proteins by computational design. Proc. Natl. Acad. Sci. U. S. A *98*, 14274-14279.

Brunt,A.A. (1992). The general properties of potyviruses. Arch. Virol. Suppl *5*, 3-16.

Bryson,J.W., Betz,S.F., Lu,H.S., Suich,D.J., Zhou,H.X., O'Neil,K.T., and DeGrado,W.F. (1995). Protein design: a hierarchic approach. Science *270*, 935-941.

Cabrita,L.D., Dai,W., and Bottomley,S.P. (2006). A family of E. coli expression vectors for laboratory scale and high throughput soluble protein production. BMC. Biotechnol. *6*, 12.

Camins,A., Verdaguer,E., Folch,J., and Pallas,M. (2006). Involvement of calpain activation in neurodegenerative processes. CNS. Drug Rev. *12*, 135-148.

Carrington,J.C. and Dougherty,W.G. (1987a). Processing of the tobacco etch virus 49K protease requires autoproteolysis. Virology *160*, 355-362.

Carrington,J.C. and Dougherty,W.G. (1987b). Small Nuclear Inclusion Protein Encoded by a Plant Potyvirus Genome Is a Protease. J. Virol. *61*, 2540-2548.

Carrington,J.C. and Dougherty,W.G. (1988). A viral cleavage site cassette: identification of amino acid sequences required for tobacco etch virus polyprotein processing. Proc. Natl. Acad. Sci. U. S. A *85*, 3391-3395.

Chames,P., Epinat,J.C., Guillier,S., Patin,A., Lacroix,E., and Paques,F. (2005). In vivo selection of engineered homing endonucleases using double-strand break induced homologous recombination. Nucleic Acids Res. *33*, e178.

Chan,D.C. and Kim,P.S. (1998). HIV entry and its inhibition. Cell *93*, 681-684.

Chevalier,B., Turmel,M., Lemieux,C., Monnat,R.J., Jr., and Stoddard,B.L. (2003). Flexible DNA target site recognition by divergent homing endonuclease isoschizomers I-CreI and I-MsoI. J. Mol. Biol. *329*, 253-269.

Chevalier,B.S., Kortemme,T., Chadsey,M.S., Baker,D., Monnat,R.J., and Stoddard,B.L. (2002). Design, activity, and structure of a highly specific artificial endonuclease. Mol. Cell *10*, 895-905.

Chevalier,B.S., Monnat,R.J., Jr., and Stoddard,B.L. (2001b). The homing endonuclease I-CreI uses three metals, one of which is shared between the two active sites. Nat Struct Biol *8*, 312-6.

Chevalier,B.S., Monnat,R.J., Jr., and Stoddard,B.L. (2001a). The homing endonuclease I-CreI uses three metals, one of which is shared between the two active sites. Nat. Struct. Biol. *8*, 312-316.

Chevalier,B.S. and Stoddard,B.L. (2001). Homing endonucleases: structural and functional insight into the catalysts of intron/intein mobility. Nucleic Acids Res *29*, 3757-74.

Choulika,A., Perrin,A., Dujon,B., and Nicolas,J.F. (1995). Induction of homologous recombination in mammalian chromosomes by using the I-SceI system of Saccharomyces cerevisiae. Mol Cell Biol *15*, 1968-73.

Christen,M., Hunenberger,P.H., Bakowies,D., Baron,R., Burgi,R., Geerke,D.P., Heinz,T.N., Kastenholz,M.A., Krautler,V., Oostenbrink,C., Peter,C., Trzesniak,D., and van Gunsteren,W.F. (2005). The GROMOS software for biomolecular simulation: GROMOS05. J. Comput. Chem. *26*, 1719-1751.

Christensen,D.D. (2007). Alzheimer's disease: progress in the development of anti-amyloid disease-modifying therapies. CNS. Spectr. *12*, 113-123.

Clarke,N.D. and Yuan,S.M. (1995). Metal search: a computer program that helps design tetrahedral metal-binding sites. Proteins *23*, 256-263.

Colonna-Cesari,F. and Sander,C. (1990). Excluded volume approximation to protein-solvent interaction. The solvent contact model. Biophys. J. *57*, 1103-1107.

Counis,M.F. and Torriglia,A. (2006). Acid DNases and their interest among apoptotic endonucleases. Biochimie *88*, 1851-1858.

Coussens,L.M., Tinkle,C.L., Hanahan,D., and Werb,Z. (2000). MMP-9 supplied by bone marrow-derived cells contributes to skin carcinogenesis. Cell *103*, 481-490.

Dahiyat,B.I., Gordon,D.B., and Mayo,S.L. (1997a). Automated design of the surface positions of protein helices. Protein Sci. *6*, 1333-1337.

Dahiyat,B.I. and Mayo,S.L. (1997a). De novo protein design: fully automated sequence selection. Science *278*, 82-87.

Dahiyat,B.I. and Mayo,S.L. (1997b). Probing the role of packing specificity in protein design. Proc. Natl. Acad. Sci. U. S. A *94*, 10172-10177.

Dahiyat,B.I., Sarisky,C.A., and Mayo,S.L. (1997b). De novo protein design: towards fully automated sequence selection. J. Mol. Biol. *273*, 789-796.

Dantas,G., Kuhlman,B., Callender,D., Wong,M., and Baker,D. (2003). A large scale test of computational protein design: folding and stability of nine completely redesigned globular proteins. J. Mol. Biol. *332*, 449-460.

Desjarlais,J.R. and Clarke,N.D. (1998). Computer search algorithms in protein modification and design. Curr. Opin. Struct. Biol. *8*, 471-475.

Desjarlais,J.R. and Handel,T.M. (1995). De novo design of the hydrophobic cores of proteins. Protein Sci. *4*, 2006-2018.

Di,V.B., Lemerle,C., Michalodimitrakis,K., and Serrano,L. (2006). From in vivo to in silico biology and back. Nature *443*, 527-533.

Donoho,G., Jasin,M., and Berg,P. (1998). Analysis of gene targeting and intrachromosomal homologous recombination stimulated by genomic double-strand breaks in mouse embryonic stem cells. Mol Cell Biol *18*, 4070-8.

Dougherty,W.G., Carrington,J.C., Cary,S.M., and Parks,T.D. (1988). Biochemical and mutational analysis of a plant virus polyprotein cleavage site. EMBO J. *7*, 1281-1287.

Dougherty,W.G., Cary,S.M., and Parks,T.D. (1989a). Molecular genetic analysis of a plant virus polyprotein cleavage site: a model. Virology *171*, 356-364.

Dougherty,W.G., Parks,T.D., Cary,S.M., Bazan,J.F., and Fletterick,R.J. (1989b). Characterization of the catalytic residues of the tobacco etch virus 49-kDa proteinase. Virology *172*, 302-310.

Doyon,J.B., Pattanayak,V., Meyer,C.B., and Liu,D.R. (2006). Directed evolution and substrate specificity profile of homing endonuclease I-SceI. J Am Chem Soc *128*, 2477-84.

Drexler,K.E. (1981). Molecular engineering: An approach to the development of general capabilities for molecular manipulation. Proc. Natl. Acad. Sci. U. S. A *78*, 5275-5278.

Dujon,B., Belfort,M., Butow,R.A., Jacq,C., Lemieux,C., Perlman,P.S., and Vogt,V.M. (1989). Mobile introns: definition of terms and recommended nomenclature. Gene *82*, 115-118.

Dunbrack,R.L., Jr. and Cohen,F.E. (1997). Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci. *6*, 1661-1681.

Durai,S., Mani,M., Kandavelou,K., Wu,J., Porteus,M.H., and Chandrasegaran,S. (2005). Zinc-Finger Nucleases: custom-designed molecular scissors for genome engineering of plant and mammalian cells. Nucleic Acids Res *33*, 5978-5990.

Eckert,D.M. and Kim,P.S. (2001). Mechanisms of viral membrane fusion and its inhibition. Annu. Rev. Biochem. *70*, 777-810.

Eisenmesser,E.Z., Kapust,R.B., Nawrocki,J.P., Mazzulla,M.J., Pannell,L.K., Waugh,D.S., and Byrd,R.A. (2000). Expression, purification, refolding, and characterization of recombinant human interleukin-13: utilization of intracellular processing. Protein Expr. Purif. *20*, 186-195.

Epinat,J.C., Arnould,S., Chames,P., Rochaix,P., Desfontaines,D., Puzin,C., Patin,A., Zanghellini,A., Paques,F., and Lacroix,E. (2003). A novel engineered meganuclease induces homologous recombination in yeast and mammalian cells. Nucleic Acids Res. *31*, 2952-2962.

Eser,M., Henrichs,T., Boyd,D., and Ehrmann,M. (2007). Target-directed proteolysis in vivo. Methods Enzymol. *421*, 68-83.

Feig,M. and Brooks,C.L., III (2004). Recent advances in the development and application of implicit solvent models in biomolecule simulations. Curr. Opin. Struct. Biol. *14*, 217-224.

Feig,M., Im,W., and Brooks,C.L., III (2004). Implicit solvation based on generalized Born theory in different dielectric environments. J. Chem. Phys. *120*, 903-911.

Fernandez-Ballester,G. and Serrano,L. (2006). Prediction of Protein Interaction Based on Structure. In Protein Design: Methods and Applications. Methods in Molecular Biology series, E.R.Guerois and M.López de la Paz, ed., pp. 207-234.

Fernandez-Ballester,G., Blanes-Mira,C., and Serrano,L. (2004). The tryptophan switch: changing ligand-binding specificity from type I to type II in SH3 domains. J. Mol. Biol. *335*, 619-629.

Fersht,A. and Winter,G. (1992). Protein engineering. Trends Biochem. Sci. *17*, 292-295.

Filikov,A.V., Hayes,R.J., Luo,P., Stark,D.M., Chan,C., Kundu,A., and Dahiyat,B.I. (2002). Computational stabilization of human growth hormone. Protein Sci. *11*, 1452-1461.

Gilis,D. and Rooman,M. (1997). Predicting protein stability changes upon mutation using database-derived potentials: solvent accessibility determines the importance of local versus non-local interactions along the sequence. J. Mol. Biol. *272*, 276-290.

Gimble,F.S., Moure,C.M., and Posey,K.L. (2003). Assessing the plasticity of DNA target site recognition of the PI-SceI homing endonuclease using a bacterial two-hybrid selection system. J. Mol. Biol. *334*, 993-1008.

Ginalski,K. (2006). Comparative modeling for protein structure prediction. Curr. Opin. Struct. Biol. *16*, 172-177.

Giver,L., Gershenson,A., Freskgard,P.O., and Arnold,F.H. (1998). Directed evolution of a thermostable esterase. Proc. Natl. Acad. Sci. U. S. A *95*, 12809-12813.

Gorbalenya,A.E., Donchenko,A.P., Blinov,V.M., and Koonin,E.V. (1989). Cysteine proteases of positive strand RNA viruses and chymotrypsin-like serine proteases. A distinct protein superfamily with a common structural fold. FEBS Lett. *243*, 103-114.

Gordon,D.B., Marshall,S.A., and Mayo,S.L. (1999). Energy functions for protein design. Curr. Opin. Struct. Biol. *9*, 509-513.

Gosert,R., Dollenmaier,G., and Weitz,M. (1997). Identification of active-site residues in protease 3C of hepatitis A virus by site-directed mutagenesis. J. Virol. *71*, 3062-3068.

Gouble,A., Smith,J., Bruneau,S., Perez,C., Guyot,V., Cabaniols,J.P., Leduc,S., Fiette,L., Ave,P., Micheau,B., Duchateau,P., and Paques,F. (2006). Efficient in toto targeted recombination in mouse liver by meganuclease-induced double-strand break. J. Gene Med. *8*, 616-622.

Guerois,R., Nielsen,J.E., and Serrano,L. (2002). Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J. Mol. Biol. *320*, 369-387.

Guerois,R. and Serrano,L. (2000). The SH3-fold family: experimental evidence and prediction of variations in the folding pathways. J. Mol. Biol. *304*, 967-982.

Gutierrez-Fernandez,A., Inada,M., Balbin,M., Fueyo,A., Pitiot,A.S., Astudillo,A., Hirose,K., Hirata,M., Shapiro,S.D., Noel,A., Werb,Z., Krane,S.M., Lopez-Otin,C., and Puente,X.S. (2007). Increased inflammation delays wound healing in mice deficient in collagenase-2 (MMP-8). FASEB J.

Harayama,S. (1998). Artificial evolution by DNA shuffling. Trends Biotechnol. *16*, 76-82.

Harbury,P.B., Plecs,J.J., Tidor,B., Alber,T., and Kim,P.S. (1998). High-resolution protein design with backbone freedom. Science *282*, 1462-1467.

Haspel,J., Blanco,C., Jacob,J., and Grumet,M. (2001). System for cleavable Fc fusion proteins using tobacco etch virus (TEV) protease. Biotechniques *30*, 60-66.

Havranek,J.J. and Harbury,P.B. (2003). Automated design of specificity in molecular recognition. Nat. Struct. Biol. *10*, 45-52.

Head-Gordon,T. and Brown,S. (2003). Minimalist models for protein folding and design. Curr. Opin. Struct. Biol. *13*, 160-167.

Heath,P.J., Stephens,K.M., Monnat,R.J., Jr., and Stoddard,B.L. (1997). The structure of I-CreI, a group I intron-encoded homing endonuclease. Nat. Struct. Biol. *4*, 468-476.

Hellinga,H.W. (1997). Rational protein design: combining theory and experiment. Proc. Natl. Acad. Sci. U. S. A *94*, 10015-10017.

Hellinga,H.W. and Richards,F.M. (1991). Construction of new ligand binding sites in proteins of known structure. I. Computer-aided modeling of sites with pre-defined geometry. J. Mol. Biol. *222*, 763-785.

Hui,K.Y., Manetta,J.V., Gygi,T., Bowdon,B.J., Keith,K.A., Shannon,W.M., and Lai,M.H. (1991). A rational approach in the search for potent inhibitors against HIV proteinase. FASEB J. *5*, 2606-2610.

Ichiyanagi,K., Ishino,Y., Ariyoshi,M., Komori,K., and Morikawa,K. (2000). Crystal structure of an archaeal intein-encoded homing endonuclease PI-PfuI. J Mol Biol *300*, 889-901.

Iqbalsyah,T.M. and Doig,A.J. (2005). Pairwise coupling in an Arg-Phe-Met triplet stabilizes alpha-helical peptide via shared rotamer preferences. J. Am. Chem. Soc. *127*, 5002-5003.

Janardhan,A. and Vajda,S. (1998). Selecting near-native conformations in homology modeling: the role of molecular mechanics and solvation terms. Protein Sci. *7*, 1772-1780.

Janin,J. and Wodak,S. (1978). Conformation of amino acid side-chains in proteins. J. Mol. Biol. *125*, 357-386.

Jiang,L., Kuhlman,B., Kortemme,T., and Baker,D. (2005). A "solvated rotamer" approach to modeling water-mediated hydrogen bonds at protein-protein interfaces. Proteins *58*, 893-904.

Joachimiak,L.A., Kortemme,T., Stoddard,B.L., and Baker,D. (2006). Computational design of a new hydrogen bond network and at least a 300-fold specificity switch at a protein-protein interface. J. Mol. Biol. *361*, 195-208.

Jorgensen,W.L., Maxwell,D.S., and Tirado-Rives,J. (1996). Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. J. Am. Chem. Soc. *118*, 11225-11236.

Jung,S., Honegger,A., and Pluckthun,A. (1999). Selection for improved protein stability by phage display. J. Mol. Biol. *294*, 163-180.

Jurica,M.S., Monnat,R.J., Jr., and Stoddard,B.L. (1998). DNA recognition and cleavage by the LAGLIDADG homing endonuclease I-CreI. Mol. Cell *2*, 469-476.

Kaplan,J. and DeGrado,W.F. (2004). De novo design of catalytic proteins. Proc. Natl. Acad. Sci. U. S. A *101*, 11566-11570.

Kapust,R.B., Tozser,J., Copeland,T.D., and Waugh,D.S. (2002). The P1' specificity of tobacco etch virus protease. Biochem. Biophys. Res. Commun. *294*, 949-955.

Kapust,R.B., Tozser,J., Fox,J.D., Anderson,D.E., Cherry,S., Copeland,T.D., and Waugh,D.S. (2001). Tobacco etch virus protease: mechanism of autolysis and

rational design of stable mutants with wild-type catalytic proficiency. Protein Eng *14*, 993-1000.

Kapust,R.B. and Waugh,D.S. (2000). Controlled intracellular processing of fusion proteins by TEV protease. Protein Expr. Purif. *19*, 312-318.

Kempkens,O., Medina,E., Fernandez-Ballester,G., Ozuyaman,S., Le,B.A., Serrano,L., and Knust,E. (2006). Computer modelling in combination with in vitro studies reveals similar binding affinities of Drosophila Crumbs for the PDZ domains of Stardust and DmPar-6. Eur. J. Cell Biol. *85*, 753-767.

Kiel,C., Serrano,L., and Herrmann,C. (2004). A detailed thermodynamic analysis of ras/effector complex interfaces. J. Mol. Biol. *340*, 1039-1058.

Kiel,C., Wohlgemuth,S., Rousseau,F., Schymkowitz,J., Ferkinghoff-Borg,J., Wittinghofer,F., and Serrano,L. (2005). Recognizing and defining true Ras binding domains II: in silico prediction based on homology modelling and energy calculations. J. Mol. Biol. *348*, 759-775.

Kim,A.R., Doherty-Kirby,A., Lajoie,G., Rylett,R.J., and Shilton,B.H. (2005). Two methods for large-scale purification of recombinant human choline acetyltransferase. Protein Expr. Purif. *40*, 107-117.

Knuesel,M., Wan,Y., Xiao,Z., Holinger,E., Lowe,N., Wang,W., and Liu,X. (2003). Identification of novel protein-protein interactions using a versatile mammalian tandem affinity purification expression system. Mol. Cell Proteomics. *2*, 1225-1233.

Koehl,P. and Levitt,M. (1999). De novo protein design. I. In search of stability and specificity. J. Mol. Biol. *293*, 1161-1181.

Kolsch,V., Seher,T., Fernandez-Ballester,G.J., Serrano,L., and Leptin,M. (2007). Control of Drosophila gastrulation by apical localization of adherens junctions and RhoGEF2. Science *315*, 384-386.

Kortemme,T., Joachimiak,L.A., Bullock,A.N., Schuler,A.D., Stoddard,B.L., and Baker,D. (2004). Computational redesign of protein-protein interaction specificity. Nat. Struct. Mol. Biol. *11*, 371-379.

Kuhlman,B., Dantas,G., Ireton,G.C., Varani,G., Stoddard,B.L., and Baker,D. (2003). Design of a novel globular protein fold with atomic-level accuracy. Science *302*, 1364-1368.

Kurtzman,A.L., Govindarajan,S., Vahle,K., Jones,J.T., Heinrichs,V., and Patten,P.A. (2001). Advances in directed protein evolution by recursive genetic

recombination: applications to therapeutic proteins. Curr. Opin. Biotechnol. *12*, 361-370.

Lawson,M.A. and Semler,B.L. (1991). Poliovirus thiol proteinase 3C can utilize a serine nucleophile within the putative catalytic triad. Proc. Natl. Acad. Sci. U. S. A *88*, 9919-9923.

Lazar,G.A., Marshall,S.A., Plecs,J.J., Mayo,S.L., and Desjarlais,J.R. (2003). Designing proteins for therapeutic applications. Curr. Opin. Struct. Biol. *13*, 513-518.

Lazaridis,T. and Karplus,M. (1999). Effective energy function for proteins in solution. Proteins *35*, 133-152.

Lazaridis,T. and Karplus,M. (2000). Effective energy functions for protein structure prediction. Curr. Opin. Struct. Biol. *10*, 139-145.

Lee,B. and Richards,F.M. (1971). The interpretation of protein structures: estimation of static accessibility. J. Mol. Biol. *55*, 379-400.

Lee,K.K., Fitch,C.A., and Garcia-Moreno,E.B. (2002). Distance dependence and salt sensitivity of pairwise, coulombic interactions in a protein. Protein Sci. *11*, 1004-1016.

Lee,M.S., Salsbury,F.R., Jr., and Olson,M.A. (2004). An efficient hybrid explicit/implicit solvent method for biomolecular simulations. J. Comput. Chem. *25*, 1967-1978.

Looger,L.L., Dwyer,M.A., Smith,J.J., and Hellinga,H.W. (2003). Computational design of receptor and sensor proteins with novel functions. Nature *423*, 185-190.

Lopez-Otin,C. and Overall,C.M. (2002). Protease degradomics: a new challenge for proteomics. Nat. Rev. Mol. Cell Biol. *3*, 509-519.

Luo,P., Hayes,R.J., Chan,C., Stark,D.M., Hwang,M.Y., Jacinto,J.M., Juvvadi,P., Chung,H.S., Kundu,A., Ary,M.L., and Dahiyat,B.I. (2002). Development of a cytokine analog with enhanced stability using computational ultrahigh throughput screening. Protein Sci. *11*, 1218-1226.

MacGregor,R.R., Chu,L.L., and Cohn,D.V. (1976). Conversion of proparathyroid hormone to parathyroid hormone by a particulate enzyme of the parathyroid gland. J. Biol. Chem. *251*, 6711-6716.

Mackerell,A.D., Jr. (2004). Empirical force fields for biological macromolecules: overview and issues. J. Comput. Chem. *25*, 1584-1604.

Mackerell,A.D., Jr., Banavali,N., and Foloppe,N. (2000). Development and current status of the CHARMM force field for nucleic acids. Biopolymers *56*, 257-265.

Maizel,J.V. (2000). SDS polyacrylamide gel electrophoresis. Trends Biochem. Sci. *25*, 590-592.

Malakauskas,S.M. and Mayo,S.L. (1998). Design, structure and stability of a hyperthermophilic protein variant. Nat. Struct. Biol. *5*, 470-475.

Marko,A.C., Stafford,K., and Wymore,T. (2007). Stochastic Pairwise Alignments and Scoring Methods for Comparative Protein Structure Modeling. J. Chem. Inf. Model.

Marshall,S.A., Lazar,G.A., Chirino,A.J., and Desjarlais,J.R. (2003). Rational design and engineering of therapeutic proteins. Drug Discov. Today *8*, 212-221.

Martin,L., Stricher,F., Misse,D., Sironi,F., Pugniere,M., Barthe,P., Prado-Gotor,R., Freulon,I., Magne,X., Roumestand,C., Menez,A., Lusso,P., Veas,F., and Vita,C. (2003). Rational design of a CD4 mimic that inhibits HIV-1 entry and exposes cryptic neutralization epitopes. Nat. Biotechnol. *21*, 71-76.

Matthews,B.W. (1995). Studies on protein stability with T4 lysozyme. Adv. Protein Chem. *46*, 249-278.

Matthews,D.A., Dragovich,P.S., Webber,S.E., Fuhrman,S.A., Patick,A.K., Zalman,L.S., Hendrickson,T.F., Love,R.A., Prins,T.J., Marakovits,J.T., Zhou,R., Tikhe,J., Ford,C.E., Meador,J.W., Ferre,R.A., Brown,E.L., Binford,S.L., Brothers,M.A., DeLisle,D.M., and Worland,S.T. (1999). Structure-assisted design of mechanism-based irreversible inhibitors of human rhinovirus 3C protease with potent antiviral activity against multiple rhinovirus serotypes. Proc. Natl. Acad. Sci. U. S. A *96*, 11000-11007.

Maymon,E., Romero,R., Pacora,P., Gervasi,M.T., Bianco,K., Ghezzi,F., and Yoon,B.H. (2000). Evidence for the participation of interstitial collagenase (matrix metalloproteinase 1) in preterm premature rupture of membranes. American Journal of Obstetrics and Gynecology *183*, 914-920.

Mayo,S.L., Olafson,B.D., and Goddard,W.A. (1990). DREIDING: a generic force field for molecular simulations. J. Phys. Chem. *94*, 8897-8909.

Mehta,R., Keawwattana,W., and Kyu,T. (2004). Growth dynamics of isotactic polypropylene single crystals during isothermal crystallization from a miscible polymeric solvent. J. Chem. Phys. *120*, 4024-4031.

Melcher,K. (2000). A modular set of prokaryotic and eukaryotic expression vectors. Anal. Biochem. *277*, 109-120.

Mendes,J., Guerois,R., and Serrano,L. (2002). Energy estimation in protein design. Curr. Opin. Struct. Biol. *12*, 441-446.

Mitsui,K., Doi,H., and Nukina,N. (2006). Proteomics of polyglutamine aggregates. Methods Enzymol. *412*, 63-76.

Mohammed,F.F., Smookler,D.S., Taylor,S.E., Fingleton,B., Kassiri,Z., Sanchez,O.H., English,J.L., Matrisian,L.M., Au,B., Yeh,W.C., and Khokha,R. (2004). Abnormal TNF activity in Timp3-/- mice leads to chronic hepatic inflammation and failure of liver regeneration. Nat. Genet. *36*, 969-977.

Mohanty,A.K., Simmons,C.R., and Wiener,M.C. (2003). Inhibition of tobacco etch virus protease activity by detergents. Protein Expr. Purif. *27*, 109-114.

Morozov,A.V. and Kortemme,T. (2005). Potential functions for hydrogen bonds in protein structure prediction and design. Adv. Protein Chem. *72*, 1-38.

Moure,C.M., Gimble,F.S., and Quiocho,F.A. (2002). Crystal structure of the intein homing endonuclease PI-SceI bound to its recognition sequence. Nat Struct Biol *9*, 764-70.

Moure,C.M., Gimble,F.S., and Quiocho,F.A. (2003). The crystal structure of the gene targeting homing endonuclease I-SceI reveals the origins of its target site specificity. J. Mol. Biol. *334*, 685-695.

Mukhopadhyay,D. and Riezman,H. (2007). Proteasome-independent functions of ubiquitin in endocytosis and signaling. Science *315*, 201-205.

Munoz,V. and Serrano,L. (1994a). Elucidating the folding problem of helical peptides using empirical parameters. Nat. Struct. Biol. *1*, 399-409.

Munoz,V. and Serrano,L. (1994b). Intrinsic secondary structure propensities of the amino acids, using statistical phi-psi matrices: comparison with experimental scales. Proteins *20*, 301-311.

Musi,V., Birdsall,B., Fernandez-Ballester,G., Guerrini,R., Salvatori,S., Serrano,L., and Pastore,A. (2006). New approaches to high-throughput structure characterization of SH3 complexes: the example of Myosin-3 and Myosin-5 SH3 domains from S. cerevisiae. Protein Sci. *15*, 795-807.

Nakayama,H., Shimamura,T., Imagawa,T., Shirai,N., Itoh,T., Sako,Y., Miyano,M., Sakuraba,H., Ohshima,T., Nomura,N., and Tsuge,H. (2006). Structure of a

Hyperthermophilic Archaeal Homing Endonuclease, I-Tsp061I: Contribution of Cross-domain Polar Networks to Thermostability. J Mol Biol.

Nallamsetty,S., Kapust,R.B., Tozser,J., Cherry,S., Tropea,J.E., Copeland,T.D., and Waugh,D.S. (2004). Efficient site-specific processing of fusion proteins by tobacco vein mottling virus protease in vivo and in vitro. Protein Expr. Purif. *38*, 108-115.

Nauli,S., Kuhlman,B., and Baker,D. (2001). Computer-based redesign of a protein folding pathway. Nat. Struct. Biol. *8*, 602-605.

Nunn,C.M., Jeeves,M., Cliff,M.J., Urquhart,G.T., George,R.R., Chao,L.H., Tscuchia,Y., and Djordjevic,S. (2005). Crystal structure of tobacco etch virus protease shows the protein C terminus bound within the active site. J. Mol. Biol. *350*, 145-155.

Overall,C.M., Tam,E.M., Kappelhoff,R., Connor,A., Ewart,T., Morrison,C.J., Puente,X., Lopez-Otin,C., and Seth,A. (2004). Protease degradomics: mass spectrometry discovery of protease substrates and the CLIP-CHIP, a dedicated DNA microarray of all human proteases and inhibitors. Biol. Chem. *385*, 493-504.

Pabo,C. (1983). Molecular technology. Designing proteins and peptides. Nature *301*, 200.

Palmer,A.E., Giacomello,M., Kortemme,T., Hires,S.A., Lev-Ram,V., Baker,D., and Tsien,R.Y. (2006). Ca2+ indicators based on computationally redesigned calmodulin-peptide pairs. Chem. Biol. *13*, 521-530.

Pantoliano,M.W., Whitlow,M., Wood,J.F., Dodd,S.W., Hardman,K.D., Rollence,M.L., and Bryan,P.N. (1989). Large increases in general stability for subtilisin BPN' through incremental changes in the free energy of unfolding. Biochemistry *28*, 7205-7213.

Paques,F. and Duchateau,P. (2007). Meganucleases and DNA double-strand break-induced recombination: perspectives for gene therapy. Curr. Gene Ther. *7*, 49-66.

Parks,T.D., Howard,E.D., Wolpert,T.J., Arp,D.J., and Dougherty,W.G. (1995). Expression and purification of a recombinant tobacco etch virus NIa proteinase: biochemical analyses of the full-length and a naturally occurring truncated proteinase form. Virology *210*, 194-201.

Parks,T.D., Leuther,K.K., Howard,E.D., Johnston,S.A., and Dougherty,W.G. (1994). Release of proteins and peptides from fusion proteins using a recombinant plant virus proteinase. Anal. Biochem. *216*, 413-417.

Petukhov,M., Cregut,D., Soares,C.M., and Serrano,L. (1999). Local water bridges and protein conformational stability. Protein Sci. *8*, 1982-1989.

Phan,J., Zdanov,A., Evdokimov,A.G., Tropea,J.E., Peters,H.K., III, Kapust,R.B., Li,M., Wlodawer,A., and Waugh,D.S. (2002). Structural basis for the substrate specificity of tobacco etch virus protease. J. Biol. Chem. *277*, 50564-50572.

Pokala,N. and Handel,T.M. (2005). Energy functions for protein design: adjustment with protein-protein complex affinities, models for the unfolded state, and negative design of solubility and specificity. J. Mol. Biol. *347*, 203-227.

Ponder,J.W. and Case,D.A. (2003). Force fields for protein simulations. Adv. Protein Chem. *66*, 27-85.

Porteus,M.H. (2006). Mammalian gene targeting with designed zinc finger nucleases. Mol Ther *13*, 438-46.

Porteus,M.H. and Baltimore,D. (2003). Chimeric nucleases stimulate gene targeting in human cells. Science *300*, 763.

Porteus,M.H. and Carroll,D. (2005). Gene targeting using zinc finger nucleases. Nat Biotechnol *23*, 967-973.

Presta,L.G. (2006). Engineering of therapeutic antibodies to minimize immunogenicity and optimize function. Adv. Drug Deliv. Rev. *58*, 640-656.

Puchta,H., Dujon,B., and Hohn,B. (1996). Two different but related mechanisms are used in plants for the repair of genomic double-strand breaks by homologous recombination. Proc Natl Acad Sci U S A *93*, 5055-60.

Puente,X.S., Sanchez,L.M., Gutierrez-Fernandez,A., Velasco,G., and Lopez-Otin,C. (2005). A genomic view of the complexity of mammalian proteolytic systems. Biochem. Soc. Trans. *33*, 331-334.

Puente,X.S., Sanchez,L.M., Overall,C.M., and Lopez-Otin,C. (2003). Human and mouse proteases: a comparative genomic approach. Nat. Rev. Genet. *4*, 544-558.

Puig,O., Caspary,F., Rigaut,G., Rutz,B., Bouveret,E., Bragado-Nilsson,E., Wilm,M., and Seraphin,B. (2001). The tandem affinity purification (TAP) method: a general procedure of protein complex purification. Methods *24*, 218-229.

Rawlings,N.D., Morton,F.R., and Barrett,A.J. (2006). MEROPS: the peptidase database. Nucleic Acids Res. *34*, D270-D272.

Reina,J., Lacroix,E., Hobson,S.D., Fernandez-Ballester,G., Rybin,V., Schwab,M.S., Serrano,L., and Gonzalez,C. (2002). Computer-aided design of a PDZ domain to recognize new target sequences. Nat. Struct. Biol. *9*, 621-627.

Reumers,J., Maurer-Stroh,S., Schymkowitz,J., and Rousseau,F. (2006). SNPeffect v2.0: a new step in investigating the molecular phenotypic effects of human non-synonymous SNPs. Bioinformatics. *22*, 2183-2185.

Rigaut,G., Shevchenko,A., Rutz,B., Wilm,M., Mann,M., and Seraphin,B. (1999). A generic protein purification method for protein complex characterization and proteome exploration. Nat. Biotechnol. *17*, 1030-1032.

Rogers,J.P., Luginbuhl,P., Shen,G.S., McCabe,R.T., Stevens,R.C., and Wemmer,D.E. (1999). NMR solution structure of alpha-conotoxin ImI and comparison to other conotoxins specific for neuronal nicotinic acetylcholine receptors. Biochemistry *38*, 3874-3882.

Ross,C.A. and Poirier,M.A. (2004). Protein aggregation and neurodegenerative disease. Nat. Med. *10 Suppl*, S10-S17.

Rouet,P., Smih,F., and Jasin,M. (1994). Introduction of double-strand breaks into the genome of mouse cells by expression of a rare-cutting endonuclease. Mol Cell Biol *14*, 8096-106.

Ryan,M.D. and Flint,M. (1997). Virus-encoded proteinases of the picornavirus super-group. J. Gen. Virol. *78 (Pt 4)*, 699-723.

Rzychon,M., Chmiel,D., and Stec-Niemczyk,J. (2004). Modes of inhibition of cysteine proteases. Acta Biochim. Pol. *51*, 861-873.

Schymkowitz,J., Borg,J., Stricher,F., Nys,R., Rousseau,F., and Serrano,L. (2005b). The FoldX web server: an online force field. Nucleic Acids Res. *33*, W382-W388.

Schymkowitz,J., Borg,J., Stricher,F., Nys,R., Rousseau,F., and Serrano,L. (2005a). The FoldX web server: an online force field. Nucleic Acids Res. *33*, W382-W388.

Schymkowitz,J.W., Rousseau,F., Martins,I.C., Ferkinghoff-Borg,J., Stricher,F., and Serrano,L. (2005c). Prediction of water and metal binding sites and their affinities by using the Fold-X force field. Proc. Natl. Acad. Sci. U. S. A *102*, 10147-10152.

Seligman,L.M., Chisholm,K.M., Chevalier,B.S., Chadsey,M.S., Edwards,S.T., Savage,J.H., and Veillet,A.L. (2002). Mutations altering the cleavage specificity of a homing endonuclease. Nucleic Acids Res. *30*, 3870-3879.

Selzer,T., Albeck,S., and Schreiber,G. (2000). Rational design of faster associating and tighter binding protein complexes. Nat. Struct. Biol. *7*, 537-541.

Serrano,L., Kellis,J.T., Jr., Cann,P., Matouschek,A., and Fersht,A.R. (1992). The folding of an enzyme. II. Substructure of barnase and the contribution of different interactions to protein stability. J. Mol. Biol. *224*, 783-804.

Shapovalov,M.V. and Dunbrack,R.L., Jr. (2007). Statistical and conformational analysis of the electron density of protein side chains. Proteins *66*, 279-303.

Shifman,J.M. and Mayo,S.L. (2003). Exploring the origins of binding specificity through the computational redesign of calmodulin. Proc. Natl. Acad. Sci. U. S. A *100*, 13274-13279.

Shifman,J.M. and Mayo,S.L. (2002). Modulating calmodulin binding specificity through computational protein design. J. Mol. Biol. *323*, 417-423.

Shifman,M.A., Srivastava,R., Brandt,C.A., Li,T.R., White,K., and Miller,P.L. (2004). Exploring the portability of informatics capabilities from a clinical application to a bioscience application. J. Am. Med. Inform. Assoc. *11*, 294-299.

Shih,Y.P., Wu,H.C., Hu,S.M., Wang,T.F., and Wang,A.H. (2005). Self-cleavage of fusion protein in vivo using TEV protease to yield native protein. Protein Sci. *14*, 936-941.

Silva,G.H., Dalgaard,J.Z., Belfort,M., and Van,R.P. (1999). Crystal structure of the thermostable archaeal intron-encoded endonuclease I-DmoI. J. Mol. Biol. *286*, 1123-1136.

Smith,J., Grizot,S., Arnould,S., Duclert,A., Epinat,J.C., Chames,P., Prieto,J., Redondo,P., Blanco,F.J., Bravo,J., Montoya,G., Paques,F., and Duchateau,P. (2006). A combinatorial approach to create artificial homing endonucleases cleaving chosen sequences. Nucleic Acids Res. *34*, e149.

Smith,T.A. and Kohorn,B.D. (1991). Direct selection for sequences encoding proteases of known specificity. Proc. Natl. Acad. Sci. U. S. A *88*, 5159-5162.

Sorin,E.J. and Pande,V.S. (2005). Empirical force-field assessment: The interplay between backbone torsions and noncovalent term scaling. J. Comput. Chem. *26*, 682-690.

Spiegel,P.C., Chevalier,B., Sussman,D., Turmel,M., Lemieux,C., and Stoddard,B.L. (2006). The structure of I-CeuI homing endonuclease: Evolving asymmetric DNA recognition from a symmetric protein scaffold. Structure. *14*, 869-880.

Steuer,S., Pingoud,V., Pingoud,A., and Wende,W. (2004). Chimeras of the homing endonuclease PI-SceI and the homologous Candida tropicalis intein: a study to explore the possibility of exchanging DNA-binding modules to obtain highly specific endonucleases with altered specificity. Chembiochem *5*, 206-13.

Stolworthy,L.D. and Shirts,R.B. (1997). ANLIZE: a molecular mechanics force field visualization tool and its application to 18-crown-6. J. Comput. Aided Mol. Des *11*, 129-134.

Street,A.G. and Mayo,S.L. (1998). Pairwise calculation of protein solvent-accessible surface areas. Fold. Des *3*, 253-258.

Strydom,D.J. (1998). The angiogenins. Cell Mol. Life Sci. *54*, 811-824.

Sussman,D., Chadsey,M., Fauce,S., Engel,A., Bruett,A., Monnat,R., Jr., Stoddard,B.L., and Seligman,L.M. (2004). Isolation and characterization of new homing endonuclease specificities at individual target site positions. J. Mol. Biol. *342*, 31-41.

Terman,A., Kurz,T., Gustafsson,B., and Brunk,U.T. (2006). Lysosomal labilization. IUBMB. Life *58*, 531-539.

Thierry,A. and Dujon,B. (1992). Nested chromosomal fragmentation in yeast using the meganuclease I-Sce I: a new method for physical mapping of eukaryotic genomes. Nucleic Acids Res *20*, 5625-31.

Topham,C.M., Srinivasan,N., and Blundell,T.L. (1997). Prediction of the stability of protein mutants based on structural environment-dependent amino acid substitution and propensity tables. Protein Eng *10*, 7-21.

Turner,B.G. and Summers,M.F. (1999). Structural biology of HIV. J. Mol. Biol. *285*, 1-32.

Urabe,M., Kume,A., Takahashi,T., Serizawa,N., Tobita,K., and Ozawa,K. (1999). A switching system regulating subcellular localization of nuclear proteins using a viral protease. Biochem. Biophys. Res. Commun. *266*, 92-96.

van den,B.B. and Eijsink,V.G. (2002). Selection of mutations for increased protein stability. Curr. Opin. Biotechnol. *13*, 333-337.

van den,B.B., Vriend,G., Veltman,O.R., Venema,G., and Eijsink,V.G. (1998). Engineering an enzyme to resist boiling. Proc. Natl. Acad. Sci. U. S. A *95*, 2056-2060.

van der Sloot,A.M., Mullally,M.M., Fernandez-Ballester,G., Serrano,L., and Quax,W.J. (2004). Stabilization of TRAIL, an all-beta-sheet multimeric protein, using computational redesign. Protein Eng Des Sel *17*, 673-680.

van der Sloot,A.M., Tur,V., Szegezdi,E., Mullally,M.M., Cool,R.H., Samali,A., Serrano,L., and Quax,W.J. (2006). Designed tumor necrosis factor-related apoptosis-inducing ligand variants initiating apoptosis exclusively via the DR5 receptor. Proc. Natl. Acad. Sci. U. S. A *103*, 8634-8639.

Ventura,S. and Serrano,L. (2004). Designing proteins from the inside out. Proteins *56*, 1-10.

Villanueva,J., Fernandez-Ballester,G., Querol,E., Aviles,F.X., and Serrano,L. (2003). Ligand screening by exoproteolysis and mass spectrometry in combination with computer modelling. J. Mol. Biol. *330*, 1039-1048.

Villegas,V., Viguera,A.R., Aviles,F.X., and Serrano,L. (1996). Stabilization of proteins by rational design of alpha-helix stability using helix/coil transition theory. Fold. Des *1*, 29-34.

Voigt,C.A., Gordon,D.B., and Mayo,S.L. (2000). Trading accuracy for speed: A quantitative comparison of search algorithms in protein sequence design. J. Mol. Biol. *299*, 789-803.

Wallace,A.C., Laskowski,R.A., and Thornton,J.M. (1995). LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. Protein Eng *8*, 127-134.

Wang,J., Kim,H.H., Yuan,X., and Herrin,D.L. (1997). Purification, biochemical characterization and protein-DNA interactions of the I-CreI endonuclease produced in Escherichia coli. Nucleic Acids Res *25*, 3767-76.

Waugh,D.S. (2005). Making the most of affinity tags. Trends Biotechnol. *23*, 316-320.

Wawersik,S., Evola,C., and Whitman,M. (2005). Conditional BMP inhibition in Xenopus reveals stage-specific roles for BMPs in neural and neural crest induction. Dev. Biol. *277*, 425-442.

Wehr,M.C., Laage,R., Bolz,U., Fischer,T.M., Grunewald,S., Scheek,S., Bach,A., Nave,K.A., and Rossner,M.J. (2006). Monitoring regulated protein-protein interactions using split TEV. Nat. Methods *3*, 985-993.

Wernisch,L., Hery,S., and Wodak,S.J. (2000). Automatic protein design with all atom force-fields by exact and heuristic optimization. J. Mol. Biol. *301*, 713-736.

Whitcomb,D.C. and Lowe,M.E. (2007). Human pancreatic digestive enzymes. Dig. Dis. Sci. *52*, 1-17.

Wisz,M.S. and Hellinga,H.W. (2003). An empirical model for electrostatic interactions in proteins incorporating multiple geometry-dependent dielectric constants. Proteins *51*, 360-377.

Zavrski,I., Kleeberg,L., Kaiser,M., Fleissner,C., Heider,U., Sterz,J., Jakob,C., and Sezer,O. (2007). Proteasome as an emerging therapeutic target in cancer. Curr. Pharm. Des *13*, 471-485.

Zhan,L., Chen,J.Z., and Liu,W.K. (2007). Computational study of the Trp-cage miniprotein based on the ECEPP/3 force field. Proteins *66*, 436-443.

Zollars,E.S., Marshall,S.A., and Mayo,S.L. (2006). Simple electrostatic model improves designed protein sequences. Protein Sci. *15*, 2014-2018.

UNIVERSITAS
Miguel
Hernández

VIII. APPENDIX

7.1 PUBLICATIONS

Fajardo-Sanchez Emmanuel, Stricher Francois, Paques Frédéric, Isalan Mark and Serrano Luis**. "Computer design of obligate heterodimer meganucleases allows efficient cutting of custom DNA sequences**" Paper submitted to PNAS, April 2007.

Fajardo-Sanchez Emmanuel, Stricher Francois, Paques Frédéric, Isalan Mark and Serrano Luis. "**Obligate heterodimer meganucleases and uses thereof**". European patent PCT/IB2007/000849, 1st February 2007.

García-Sanz N, Fernanández-Carvajal A, Morenilla-Palao C, Planells-Cases R, **Fajardo-Sánchez E,** Fernández-Ballester G, Ferrer-Montiel A. **"Identification of a tetramerization domain in the C terminus of the vanilloid receptor."** J.Neurosci. 2004 Jun 9;24(23):5307-14.

UNIVERSITAS
Miguel
Hernández