Bioengineering Coagulation Factor Xa Substrate Specificity into

*Streptomyces griseus* Trypsin

by

Michael J. Page

B.Sc., Carleton University, 1998

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

Department of Biochemistry and Molecular Biology

We accept this thesis as conforming
to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA
April 2004

© Michael J. Page, 2004
Abstract

Extended substrate specificity is exhibited by a number of highly evolved members of the S1 peptidase family, such as the vertebrate blood coagulation proteases. Dissection of this substrate specificity has been hindered by the complexity and physiological requirements of these proteases. In order to understand the mechanisms of extended substrate specificity, a bacterial trypsin-like enzyme, *Streptomyces griseus* trypsin (SGT), was chosen as a scaffold for the introduction of extended substrate specificity through structure-based genetic engineering.

Recombinant and mutant SGT proteases were produced in a *B. subtilis* expression system, which constitutively secretes active protease into the extracellular medium at greater than 15 mg/L of culture. Comparison of the recombinant wild-type protease to the natively produced enzyme demonstrated near identity in enzymatic and structural properties. To begin construction of a high specificity protease, four mutants in the S1 substrate binding pocket (T190A, T190S, T190V, and T190P) were produced and examined for differences in the Arg:Lys preference. Only the T190P mutant of SGT demonstrated a significant increase in P1 arginine to lysine preference – a three-fold improvement to 16:1 – with only a minor reduction in catalytic activity (k_cat reduction of 25%). The 1.9 Å resolution crystal structure of T190P mutant of SGT in complex with the small molecule inhibitor benzamidine was subsequently determined. The model shows that the increased preference for Arg over Lys side chains in the S1 pocket is the result of the second shell residues of the S1 pocket, particularly by the N-terminal residue of the protease which does not conflict with the introduced proline ring.
Using the T190P mutant of SGT as a starting point, coagulation factor Xa (FXa) substrate specificity determinants were then introduced by additional site-directed mutagenesis. To aid in purification of the recombinant proteases a hexa-histidine tag was added to the C-terminus of the protein. Addition of the purification tag reduced the ability of the expression host to produce the enzyme (3 mg/L of culture) but simplified the purification of SGT from the culture medium. Various combinations of a two-residue loop and a number of point mutations at positions 99, 172, 174, 180, and 217 were constructed and characterized in SGT. The mutant bearing mutations at all positions except residue 217 demonstrated a moderate preference for FXa substrates as determined using chromogenic synthetic peptides. However, the kinetic properties of the mutant enzyme suggested that the 172-loop, a member of the S3/S4 substrate binding pocket, is not in a conformation similar to FXa. Addition of the Y217E mutation was designed to stabilize the loop but led to a specific protease similar to coagulation factor XIa and not FXa. These results confirm the evolutionary relationship amongst the vertebrate coagulation proteases and demonstrate the importance and flexibility of the 172-loop. Further, Na\textsuperscript{+} binding, a novel property found in several coagulation proteases, is suggested to play a role in stabilization of the 172-loop and in turn played an important role in the evolution of the vertebrate coagulation cascade.
# Table of Contents

Abstract ................................................................................................................................. ii  

Table of Contents ................................................................................................................ iv  

List of Tables ....................................................................................................................... vii  

List of Figures ........................................................................................................................ viii  

List of Abbreviations .......................................................................................................... x  

Acknowledgments ................................................................................................................. xi  

Chapter 1. Introduction ....................................................................................................... 1  
  1.1 Focus of Research ........................................................................................................... 1  
  1.2 Overview ....................................................................................................................... 1  
  1.3 Nomenclature of Proteases ............................................................................................ 4  
  1.4 Definition of Substrate Specificity .................................................................................. 6  
  1.5 Aspartic, Cysteine, Metallo-, and Threonine Proteases .................................................. 7  
  1.6 Catalytic Mechanism of Serine Proteases ...................................................................... 9  
  1.7 Blood Coagulation and Fibrinolysis .............................................................................. 14  
  1.8 Substrate Specificity of Serine Proteases .................................................................... 17  
  1.9 Genetic Manipulation of Serine Proteases ................................................................... 20  
  1.10 Streptomyces griseus Trypsin ..................................................................................... 21  
  1.11 Statement of Hypothesis ............................................................................................. 22  
  1.12 Objectives and Outline ............................................................................................... 22  

Chapter 2. Recombinant Protein Expression of Streptomyces griseus Trypsin in Bacillus subtilis .......................................................................................................................... 25  
  2.1 Introduction ................................................................................................................... 25  
  2.2 Materials & Methods .................................................................................................... 28  
    2.2.1 Plasmids, Bacterial Strains, and Growth Conditions ................................................. 28  
    2.2.2 DNA manipulation .................................................................................................. 28  
    2.2.3 Protein Purification ................................................................................................. 29  
    2.2.4 Kinetic Analysis ...................................................................................................... 30  
    2.2.5 Crystallization ........................................................................................................ 30  
  2.3 Results and Discussion ................................................................................................. 31
2.3.1 Production and Purification of Recombinant SGT ........................................... 31
2.3.2 Kinetic Analysis ......................................................................................... 36
2.3.3 Crystallization and Structure Determination .............................................. 37
2.3.4 Comparison of the recombinant and Pronase-derived crystal structure ...... 40
2.3.5 Wild-type Native and Recombinant SGT Substrate Binding ...................... 41

2.4 Conclusions ................................................................................................. 42

Chapter 3. Engineering the Primary Substrate Specificity of Streptomyces griseus Trypsin ............................................................... 43
3.1 Introduction ...................................................................................................... 43

3.2 Materials & Methods .................................................................................... 44
  3.2.1 DNA Manipulation and Protein Purification .............................................. 44
  3.2.2 Kinetic Analysis ......................................................................................... 45
  3.2.3 Crystallization and Structure Determination .............................................. 45

3.3 Results and Discussion .................................................................................. 46
  3.3.1 Production of SGT Mutants ..................................................................... 46
  3.3.2 Kinetic Analysis ......................................................................................... 46
  3.3.3 Crystallization and Structure Refinement ................................................. 48
  3.3.4 Ca\(^{2+}\) binding site of B. subtilis derived SGT ........................................ 50
  3.3.5 T190S and the Loss of \(\gamma\)-CH\(_3\) ................................................................. 51
  3.3.6 T190V and the Effect of a Branched Side Chain ....................................... 53
  3.3.7 T190A and the Loss of \(\gamma\)-OH ................................................................. 53
  3.3.8 T190P ...................................................................................................... 54
  3.3.9 Second Shell Residues ............................................................................. 58

3.4 Conclusions .................................................................................................... 58

Chapter 4. Engineering Coagulation Factor Xa Substrate Specificity into Streptomyces griseus Trypsin .................................................... 59
4.1 Introduction ...................................................................................................... 59
  4.1.1 Overview ................................................................................................. 59
  4.1.2 Choice of Mutations to Mimic FXa-like Specificity .................................. 62

4.2 Materials & Methods .................................................................................... 65
  4.2.1 Plasmids, Bacterial Strains, and Growth Conditions ............................... 65
  4.2.2 Construction of a Hexahistidine-tagged SGT ............................................ 65
  4.2.3 Sequence analysis of the S1 Family peptidases ....................................... 67
  4.2.4 DNA Manipulation .................................................................................. 67
  4.2.5 Purification of His-tagged SGT and Mutants thereof ............................... 68
  4.2.6 Characterization of Substrate Specificity ................................................ 69
  4.2.7 Macromolecular Substrate Specificity ..................................................... 69

4.3 Results & Discussion ..................................................................................... 70
  4.3.1 Production of His-tagged SGT ................................................................. 70
  4.3.2 Techniques for Characterization of Substrate Specificity of Serine Proteases .......................... 71
List of Tables

Table 1.1 A short list of biological processes involving proteolysis ........................................3
Table 1.2 A short list of pathologies involving abnormal proteolysis .........................................3
Table 1.3 Select examples of successful protein engineering of serine proteases .........................21
Table 2.1 Purification table for recombinant SGT (bSGT) from B. subtilis extracellular supernatant .............................................................................................................................34
Table 2.2 P1 arginine to lysine preference of SGT enzymes ..........................................................37
Table 2.3 Data collection and refinement statistics of wild-type recombinant SGT .....................39
Table 3.1 Oligonucleotides used to mutate residue 190 in the SGT gene ....................................45
Table 3.2 ES-MS analysis of the four mutants of SGT .................................................................46
Table 3.3 P1 Arginine to Lysine preference of mutant SGT enzymes ..........................................47
Table 3.4 K_i values of benzamidine for recombinant SGT and the four mutant forms ..............48
Table 3.5 Data collection and refinement statistics for the T190P mutant of SGT .....................49
Table 4.1 Mutants of SGT constructed to mimic the substrate specificity of FXa .....................62
Table 4.2 Oligonucleotides used to mutate the SGT gene to mimic residues found in FXa ....68
Table 4.3 Steady-state kinetic parameters for the hydrolysis of a series of p-nitroanilide chromogenic substrates by the YSFMP mutant of SGT .........................................................83
Table 4.4 Steady-state kinetic parameters for the hydrolysis of a series of p-nitroanilide chromogenic substrates by the YSFMPE mutant of SGT .........................................................85
Table 4.5 Cleavage sites of proteases used in processing recombinant proteins .......................90
Table 5.1 Substrate specificities of S1 family peptidases ...........................................................96
Table 5.2 Substrate specificity determinants of S1 family sub-family A peptidases ...............100
Table 5.3 Inhibition constants of ecotin against a variety of S1 family peptidases .................104
List of Figures

Figure 1.1 Distribution of identified proteases based on catalytic type ...........................................2
Figure 1.2 Derivation of the Michaelis-Menten equation .................................................................6
Figure 1.3 Catalytic mechanism of a typical zinc metalloprotease ..................................................9
Figure 1.4 Catalytic triad of serine proteases ....................................................................................10
Figure 1.5 Three dimensional structure of a typical serine protease .............................................12
Figure 1.6 Catalytic mechanism of a serine protease .....................................................................13
Figure 1.7 Overview of vertebrate blood coagulation ....................................................................15
Figure 1.8 Protein domains of the vertebrate blood coagulation proteases ..................................16
Figure 1.9 Schecter & Berger nomenclature of protease specificity .............................................18
Figure 2.1 Plasmid map of SGT gene cloned into pWB980 .........................................................33
Figure 2.2 ES-MS spectrum of purified recombinant SGT ..............................................................35
Figure 2.3 SDS-PAGE of purified recombinant SGT .........................................................................36
Figure 2.4 Ramachandran plot of the crystal structure of recombinant wild-type SGT .............38
Figure 2.5 Superimposition of the Cα traces of native and recombinant wild-type SGT .............41
Figure 3.1 Ramachandran plot of the crystal structure of the T190P mutant of SGT .................50
Figure 3.2 Comparison of the Ca2+ binding site in SGT enzymes .................................................52
Figure 3.3 Comparison of the S1 binding pocket in SGT enzymes ...............................................56
Figure 4.1 Residues involved in the extended substrate specificity of coagulation proteases ........63
Figure 4.2 Plasmid construction for the production of recombinant His-tagged SGT .................66
Figure 4.3 Purification of a typical His-tagged mutant of SGT .......................................................71
Figure 4.4 Substrates used to characterize mutants of SGT with altered substrate specificity 73
Figure 4.5 Normalized kcat/Km values for the T190P, LP and YP mutants of SGT .................75
Figure 4.6 S3 & S4 binding pockets of FXa ....................................................................................77
Figure 4.7 S3 and S4 binding pockets of FVIIa ..............................................................................78
Figure 4.8 Conformation of the 172-loop in S1 peptidases ............................................................80
Figure 4.9 Normalized kcat/Km values for the YFP and YSFP mutants of SGT .........................82
Figure 4.10 Na+-binding site in thrombin .......................................................................................88
Figure 4.11 Prothrombin processing by mutants of SGT ..............................................................91
### List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Å</td>
<td>Angstrom unit (1 Å = 0.1nm)</td>
</tr>
<tr>
<td>AMC</td>
<td>Aminomethylcoumarin</td>
</tr>
<tr>
<td>aPC</td>
<td>Activated protein C</td>
</tr>
<tr>
<td>B</td>
<td>Crystallographic thermal factor (Å&lt;sup&gt;2&lt;/sup&gt;)</td>
</tr>
<tr>
<td>Bz</td>
<td>Benzoyl</td>
</tr>
<tr>
<td>dNTP</td>
<td>Dinucleotidetriphosphate</td>
</tr>
<tr>
<td>$Fo, Fc$</td>
<td>Observed and calculated structure factors</td>
</tr>
<tr>
<td>FX</td>
<td>Coagulation factor X (similar for other coagulation proteases)</td>
</tr>
<tr>
<td>FXa</td>
<td>Activated coagulation factor X (similar for other coagulation proteases)</td>
</tr>
<tr>
<td>$k_{cat}$</td>
<td>Catalytic constant</td>
</tr>
<tr>
<td>$K_m$</td>
<td>Michaelis constant</td>
</tr>
<tr>
<td>LB</td>
<td>Luria – Bertani broth</td>
</tr>
<tr>
<td>NMWL</td>
<td>Nominal molecular weight limit</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase Chain Reaction</td>
</tr>
<tr>
<td>Pip</td>
<td>Pippecoyl</td>
</tr>
<tr>
<td>pNA</td>
<td>para-nitroanilide</td>
</tr>
<tr>
<td>r.m.s.</td>
<td>Root mean squared</td>
</tr>
<tr>
<td>$R_{free}$</td>
<td>R factors based test set of excluded reflections</td>
</tr>
<tr>
<td>$R_{merge}$</td>
<td>$\Sigma_{hkl} \Sigma_i</td>
</tr>
<tr>
<td>$R_{cryst}$</td>
<td>$\Sigma</td>
</tr>
<tr>
<td>S.D.</td>
<td>Standard deviation at 95% confidence interval</td>
</tr>
<tr>
<td>SDS-PAGE</td>
<td>Sodium dodecyl sulfate – polyacrylamide gel electrophoresis</td>
</tr>
<tr>
<td>Tos</td>
<td>Tosyl</td>
</tr>
<tr>
<td>Tris-HCl</td>
<td>Tris(hydroxymethyl)aminomethane hydrochloride</td>
</tr>
</tbody>
</table>
Acknowledgments

I am indebted to Ross and Jeff for countless hours of good times, their use of "Number 1", and their ability to define new verbs.

I would especially like to thank Sui-Lam Wong for providing the essential components of the expression system.

I appreciate my parents for their constant support both emotionally and financially throughout my life.

I thank my friends, for their help, advice, time and patience. I especially thank Iain, Marty, Ismail, Angus, Mike K., and Mark and wish them the best in all endeavors.

A special thanks to Tanya for all of her helpful comments.

This thesis is dedicated to my beloved Doe
Chapter 1. Introduction

1.1 Focus of Research

How do you create a highly efficient and highly specific enzyme? Over several million years, nature has evolved many enzymes that possess high degrees of specificities. Using methods of genetic manipulation, properties of enzymes can be altered, including their substrate specificity. The present dissertation involves taking a primitive bacterial enzyme and adding substrate specificity where little existed previously. By doing so, we will learn about how molecular evolution has produced substrate specificity in one family of enzymes, the serine proteases. I will use a human protease involved in blood coagulation as our guide and then extend the concepts learned to a system that could produce a variety of other specificities. To begin, I will discuss what is a protease, what do they do and why they are important. From there, I will describe the system and the methods applied to generate a specific protease, what has been done by others and then by the author. Lastly, I will discuss the opportunities that may result from this work.

1.2 Overview

Proteolytic enzymes play a diverse number of roles in a variety of essential biological processes, both as non-specific catalysts of protein degradation and as highly specific agents that control physiological events. Hydrolysis of a peptide bond is the key function that proteases fulfill in vivo, and this may result in the activation or destruction of its substrate. Numerous biological processes involving proteolytic activity have been characterized and a wealth of information has been gathered on the five major catalytic classes of these enzymes (Figure 1.1, Table 1.1). Roughly 2% of all genes in most organisms are proteases, second
only to transcription factors. Hence, the importance of these types of enzymes in biological and commercial settings cannot be understated.

![Pie chart showing the distribution of identified proteases based on catalytic type.](image)

**Figure 1.1** Distribution of identified proteases based on catalytic type.

Numerous pathological conditions are the result of excessive or insufficient proteolytic activity (Table 1.2). These clinical situations can arise from genetic defects in the proteases themselves, their natural substrates or their natural inhibitors. A number of academic laboratories and pharmaceutical companies are devoted to the production of therapeutic products, such as small molecule inhibitors or recombinant proteins, to minimize the effects of protease-related pathologies [1,2]. Given the large number of closely related proteases found in man, the design of potent inhibitors with minimal cross reactivity is an arduous task. Detailed information on the active site geometry and electronic configuration of the protease is a requirement for proper inhibitor design. These studies have been hampered
### Biological Process

<table>
<thead>
<tr>
<th>Biological Process</th>
<th>Proteolytic Event</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apoptosis</td>
<td>Control of physiological cell death</td>
<td>[3]</td>
</tr>
<tr>
<td>Blood Coagulation</td>
<td>Proteolytic cascades of clot formation, fibrinolysis</td>
<td>[4]</td>
</tr>
<tr>
<td>Blood Pressure</td>
<td>Renin-angiotensin and kallikrein-kinin systems</td>
<td>[5]</td>
</tr>
<tr>
<td>Digestion</td>
<td>Breakdown of protein into tri- and dipeptides; liberation of hormones promoting digestion</td>
<td>[6]</td>
</tr>
<tr>
<td>Fertilization</td>
<td>Sperm-Egg interaction, ovulation, ovum implantation and parturition</td>
<td>[7]</td>
</tr>
<tr>
<td>Immunity</td>
<td>Complement activation, antigen presentation, chemokine and chemotaxin activation;</td>
<td>[8]</td>
</tr>
<tr>
<td>Intracellular protein level</td>
<td>Proteasome-Ubiquitin system</td>
<td>[9]</td>
</tr>
<tr>
<td>Protein Processing</td>
<td>Zymogen activation and protein sorting</td>
<td>[10]</td>
</tr>
<tr>
<td>Tissue Remodeling</td>
<td>Turnover and repair of the extracellular environment</td>
<td>[11]</td>
</tr>
</tbody>
</table>

#### Table 1.1
A short list of biological processes involving proteolysis.

### Pathology

<table>
<thead>
<tr>
<th>Pathology</th>
<th>Proteolytic Event</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alzheimer's Disease</td>
<td>Processing of amyloid precursor protein</td>
<td>[12]</td>
</tr>
<tr>
<td>Cancer</td>
<td>Regulation of apoptosis, tumor growth and invasion</td>
<td>[3]</td>
</tr>
<tr>
<td>Chronic Inflammation</td>
<td>Excessive activation of pro-inflammatory cytokines</td>
<td>[13]</td>
</tr>
<tr>
<td>Hemophilia</td>
<td>Insufficient levels of coagulation factor activity and slowed clot formation</td>
<td>[14]</td>
</tr>
<tr>
<td>Myocardial Infarction</td>
<td>Unregulated coagulation leading to restricted blood flow</td>
<td>[15]</td>
</tr>
<tr>
<td>Parasite Infection</td>
<td>Regulation of the parasitic life-cycle</td>
<td>[16]</td>
</tr>
<tr>
<td>Viral Replication</td>
<td>Processing of viral coat proteins and other essential replication machinery required for viral infection</td>
<td>[17]</td>
</tr>
</tbody>
</table>

#### Table 1.2
A short list of pathologies involving abnormal proteolysis.
by the ability to produce large quantities of the target enzyme and detailed three dimensional structural models of them.

Industrial usage of proteolytic enzymes is widespread and commercially important. For example, the production of the protease subtilisin for use in detergents is on the scale of tons per annum and accounts for 40% of enzyme sales worldwide. Other industrial applications of proteases include the production of food stuffs, leather, pharmaceuticals, diagnostic reagents, waste management, and silver recovery [18,19]. Recent advances in molecular biology technology have allowed for the development of proteases with improved properties for industrial use. Some successful examples of protein engineering include thermostability, resistance to oxidation, alteration of pH optima, and increased catalytic efficiency [20]. However, the use of highly specific proteases for site specific proteolysis has for the most part been limited to academic endeavors due to the difficulties associated with producing high purity enzymes from the complex biological systems from which they derive.

In the present dissertation, a novel system has been developed to produce a recombinant protease and the substrate specificity of the enzyme altered. The results contribute to our understanding of the substrate specificity determinants of the trypsin-like family of serine proteases. Furthermore, a framework for designing novel specificities not observed in nature is then presented based on the work.

1.3 Nomenclature of Proteases

Prior to a literature review, a few words must be mentioned on the nomenclature used to describe proteolytic enzymes. The International Union of Biochemical and Molecular Biology Nomenclature Committee (IUBMB-NC) denotes hydrolases acting on peptide bonds
as E.C.3.4, yet only a small fraction of the known proteases have been given full
classification. Given the wide distribution of proteolytic enzymes and their historical
significance, a variety of common names are used. Papain, commonly used to tenderize meat,
owes its name to the papaya plant from which it is derived. Other proteases are named based
on the physiological process in which they were first described, such as coagulation factor X
which is involved in vertebrate blood coagulation. Unfortunately, a significant number of
proteases owe their name to having similar biochemical properties yet are involved in
disparate biological processes or having different catalytic mechanisms. For example, two of
the many proteases found in the human liver, cathepsin B and cathepsin D, are named
similarly but belong to the cysteine and aspartic protease family, respectively. Thus, the
nomenclature for each particular protease has become rather muddled with convention.

A variety of terms have been used to refer to enzymes that catalyze the hydrolysis of a
peptide bond. The term peptidase has been suggested by the IUBMB-NC as a preferred
alternative to protease [21]. Peptidase refers more easily to both exopeptidases (enzymes that
cleave peptide bonds at the ends of a polypeptide chain), endopeptidases (those that cleave
peptide bonds in the middle of polypeptide chains), and oligopeptidases (those that cleave
only short polypeptides). Exopeptidases are further subdivided into aminopeptidases (cleaving
at the N-terminus) and carboxypeptidases (cleaving at the C-terminus). Other terms found in
the literature include proteinase and proteolytic enzyme. In the present thesis, the terms
protease and peptidase will be used interchangeably to describe the enzymes studied in the
research; all of which are endopeptidases.
1.4 Definition of Substrate Specificity

Enzyme kinetics are traditionally defined by Michaelis-Menten kinetics which apply a steady state assumption to the reaction (Figure 1.2). Substrate specificity is traditionally defined by the ratio of the maximum reaction velocity catalyzed by the enzyme per unit time ($k_{cat}$) to the Michaelis-Menten constant ($K_m$). $k_{cat}$ is equal to the maximal reaction velocity ($V_{max}$) divided by the total amount of enzyme in the reaction. In theory, both constants ($k_{cat}$ and $K_m$) are linked mathematically as $K_m$ is dependent on the rates of each step in the reaction ($K_m = (k_{-1} + k_2) / k_1$). In the present study, the traditional definition of substrate specificity will be applied ($S = k_{cat}/K_m$).

\[ E + S \xrightarrow{k_1} ES \xrightarrow{k_2} E + P \]

Initial Rate of Reaction: \( v = \frac{d[P]}{dt} = k \times [ES] \)
Steady state assumption: \( \frac{d[ES]}{dt} = 0 \)
Rate of ES formation = Rate of ES breakdown
\( k_1[E][S] = k_{-1}[ES] + k_2[ES] \)

Figure 1.2 Derivation of the Michaelis-Menten equation. For a typical enzyme catalyzed reaction (A) the rate of change of the enzyme (E) in the enzyme-substrate complex (ES) is assumed to be constant throughout the reaction. The Michaelis-Menten equation can then be derived (B) by introducing the Michaelis constant ($K_m$), which is only true if the rate of ES formation ($k_1$) is much larger than the rate of product formation ($k_2$).
1.5 Aspartic, Cysteine, Metallo-, and Threonine Proteases

Historically, four mechanistic classes of proteolytic enzymes have been recognized based on their catalytic mechanism – aspartic, cysteine, metallo-, and serine proteases. With the advent of whole genome sequencing this classification system has become inadequate as the variety of catalytic mechanisms identified in nature has expanded rapidly. At present, nearly 18,000 gene sequences for peptidases have been identified and over 2000 peptidases have been characterized. Barrett has devised a classification scheme based on statistically significant similarities in sequence and structure of all known proteolytic enzymes, and terms this database MEROPS (www.merops.ac.uk) [22,23]. This system divides all known proteases into 40 clans and over 169 sub-families, not including a group of putative proteases of unknown mechanism. Only the four historic classifications of proteolytic enzymes will be mentioned here.

Hydrolysis of a polypeptide backbone requires three key mechanistic hurdles to be overcome for efficient catalysis to proceed. Peptide bonds are particularly stable due to the electron resonance between the amide nitrogen and carbonyl group of the bond. By the use of a general acid, proteases overcome this partial double bond character through the generation of a negatively charged tetrahedral intermediate that is stabilized by the active site. Secondly, water is a poor nucleophile and must be activated, typically via a general base. Lastly, amines are poor leaving groups and must be expelled from the active site prior to completion of the catalytic cycle [24]. Proteases accomplish these tasks efficiently and increase the rate of reaction $\sim 10^{10}$-fold over the uncatalyzed reaction. Moreover, proteases can catalyze similar reactions in the hydrolysis of amides, esters, anilides, and thioesters.
Aspartic proteases and metalloproteases catalyze the hydrolysis of the polypeptide backbone through activation of a water molecule. Metalloproteases comprise the second largest family of proteases known in nature and typically utilize a zinc ion in their active site; however, cobalt or manganese are also found (Figure 1.3). In many metalloproteases, a single metal ion is utilized in the catalysis; however two metal ions acting cocatalytically are typically found in proteases containing cobalt or manganese. Three amino acid side chains, typically His, Glu, Asp or Lys residues, are involved in the co-ordination of the metal ion and at least one other residue is required for catalysis [25]. The catalytic residue is typically Glu in many metallopeptidases but alternatives exist. Activation of the water molecule in the aspartic family of proteases is the result a pair of Asp residues that co-ordinate the activated water molecule [26].

In contrast to nucleophilic attack of the amide backbone by an activated water molecule, cysteine, serine, and threonine peptidases utilize an amino acid side chain. Cysteine proteases compromise the third largest family of known peptidases and employ a nucleophilic sulfhydryl from a cysteine residue to catalyze the hydrolysis of an amide bond [27]. Although less abundant in nature, threonine peptidases are deeply involved in the key biological process of intracellular degradation of polypeptides by the proteasome [28]. The overall catalytic mechanism of both of these families of protease is more similar to that found in serine proteases, where a nucleophile as well as a proton donor is required for catalysis. The proton donor in all cysteine and threonine peptidases which have been identified is a His residue, which is also true of all known serine proteases.
1.6 Catalytic Mechanism of Serine Proteases

Nearly a third of all known proteases are classified in the serine protease family of enzymes. The family name stems from the nucleophilic serine residue in the active site of the enzyme, and the catalytic potency of this residue is dependent on the Asp-Ser-His charge relay system or catalytic triad which was originally proposed by Blow over 30 years ago (Figure 1.4) [29]. These three residues are found in an identical structural position in four different three-dimensional protein folds that catalyze the hydrolysis of peptide bonds,
suggested four distinct evolutionary origins. Common examples of these folds are represented by chymotrypsin, subtilisin, carboxypeptidase Y, and Clp protease. A number of other enzyme families, including asparaginases, esterases, acylases, and β-lactamases, utilize the Asp-Ser-His catalytic triad or variants to generate a strong nucleophile and promote catalysis [30]. For the remainder of the introduction, I will limit the discussion to the chymotrypsin family (S1 peptidase family) of serine proteases, which includes trypsins and elastases. Moreover, I will utilize the chymotrypsin numbering system suggested by Blow to refer to a particular amino acid residue. It must be noted that many of the concepts discussed in relation to chymotrypsin-like proteases apply similarly to other types of proteases.

![Catalytic triad of serine proteases](image)

**Figure 1.4** Catalytic triad of serine proteases. The Asp, His, and Ser combination is found in serine proteases and other enzyme families that require a nucleophilic serine side chain.

Central to the catalytic triad is the existence of a hydrogen bond between residue Asp102 and His57, which facilitates the abstraction of the proton from Ser195 and generates a potent nucleophile. Some controversy exists over whether this hydrogen bond can be described as a low barrier hydrogen bond (LBHB), an instance where the pK values between
the donor and acceptor are matched. Rejection of the LBHB theory mainly stems from the argument that it would provide no significant improvement to catalytic rate enhancement [31,32]. Increasing experimental and theoretical data are supporting this theory and the debate continues [33,34]. Stabilization of the catalytic triad is mediated through a network of additional hydrogen bonds provided by several highly conserved amino acid residues surrounding the triad, particularly Ala56 and Ser214 in the chymotrypsin family of serine proteases. Significant effort has been placed in the development of small molecule compounds that mimic the activity the catalytic triad, but have met with limited success due to the complexity of the chemistry involved to generate the nucleophilic serine.

Activation of chymotrypsin-like serine proteases requires proteolytic processing of an inactive zymogen precursor protein. This cleavage occurs at the identical position in all known members of the family: between residues 15 and 16 [24]. The newly created N-terminus produces a conformational change in the enzyme and stabilizes the oxyanion hole and substrate binding site through formation of an electrostatic interaction with Asp194 [29]. Two β-barrel domains, each formed by six anti-parallel β-strands, and a C-terminal α-helix comprise the mature form of the enzyme (Figure 1.5). Both the catalytic residues and substrate binding site lie in the cleft between the β-barrel domains, and enzyme-substrate interactions occur with both domains. A minimum of three disulphide bonds is required to stabilize the overall structure; however five or six are commonly found in the family of enzymes.
Figure 1.5 Three dimensional structure of a typical serine protease of the S1 family of peptidases. Components of the catalytic triad are shown in stick form and are located in between the two β-barrel domains.

Figure 1.6 depicts the generally accepted mechanism of serine protease catalyzed hydrolysis of a peptide bond [24]. Initially, the hydroxyl oxygen of Ser195 attacks the carbonyl of the peptide substrate as a result of His57 in the catalytic triad acting as a general base (Steps I and II). The oxyanion tetrahedral intermediate is stabilized by the backbone atoms of Gly193 (not depicted) and Ser195 that generate a positively charged pocket within the active site (Step III). Collapse of the tetrahedral intermediate generates the acyl-enzyme intermediate and stabilization of the newly created N-terminus is mediated by His57 (Steps IV and V). Evidence for the existence of the acyl-enzyme intermediate was provided in 1954 by Hartley and Kilbey [35]. In these initial experiments a pre-steady state burst of product
Figure 1.6 Catalytic mechanism of a serine protease. Formation of a tetrahedral intermediate is the key conformational step in the acylation and deacylation reactions.
correctly identified that a bond to a hydroxyl moiety within chymotrypsin was involved in the 
reaction mechanism. In the second half of the mechanism, a water molecule displaces the free 
polypeptide fragment and attacks the acyl-enzyme intermediate (Step VI). Again, the 
oxyanion hole stabilizes the second tetrahedral intermediate of the pathway and collapse of 
this intermediate liberates a new C-terminus.

1.7 Blood Coagulation and Fibrinolysis

Vertebrate blood coagulation and fibrinolysis can serve as a useful paradigm for the 
study of proteolysis in a biological setting. The process serves as a model for pathologies 
associated with improper proteolysis, molecular evolution through gene duplication and 
divergence, as well as understanding molecular recognition and substrate specificity. Initially 
recognized as a cascade of events that leads to amplification and rate enhancement, the 
feedback pathways of the clotting cascade have only recently become elucidated.

At the site of an injury that leads to disruption of the integrity of a blood vessel, a 
rapid and specific response must be employed to prevent excessive blood loss and to restrict 
bacterial infection [36]. In vivo, the formation of a fibrin clot requires a minimum of five 
proteases: coagulation factor XI (FXI), coagulation factor IX (FIX), coagulation factor VII 
(FVII), coagulation factor X (FX), and prothrombin. These enzymes circulate in the blood 
stream at low concentrations in inactive, zymogen forms. Activation of these zymogens 
requires the site-specific proteolysis of one or more peptide bonds, liberating a free N-
terminus and promoting proteolytic activity (active forms of the enzymes are denoted with a 
lower case “a”, such as FXa) [37]. Localization of the proteases to membrane surfaces at the 
site of injury is provided by three co-factors: activated coagulation factor V (FVa), activated
coagulation factor VIII (FVIIIa) and tissue factor (TF) [38]. In combination with a phospholipid bilayer these co-factors promote the rate of clot formation $\sim 10^6$-fold (Figure 1.7). Down-regulation of the pathway is mediated in part by another protease, activated protein C (aPC), which cleaves two of the co-factors, FVa and FVIIIa, at specific positions in the protein and inactivates them [39]. Thrombin plays a pivotal role in the process as it activates protein C, FV, and FVIII as well as a number of other signaling proteins that recruit cells and proteins to the site of damage [40].

**Figure 1.7** Overview of vertebrate blood coagulation. Five proteases are involved in the formation of a cross-linked fibrin blood clot (factors Xla, IXa, VIIa, Xa, and thrombin). Three accessory proteins (factors VIIIa and Va, and tissue factor (TF)) are involved in co-localization on a phospholipid surface (PL) and catalytic rate enhancement of the entire process. Protein C (PC) is activated (aPC) by thrombin and leads to inhibition of the process by cleaving the co-factors.
Biochemical characterization of the purified components of the blood coagulation pathway *in vitro* has shown that each protease in the pathway prefers to recognize and cleave a particular sequence of amino acids. Substrate specificity of these proteases, however, is not extremely strict owing to the requirement for the process to occur rapidly [41,42]. Substrate specificity is a combination of additional regions of the enzyme that contribute to molecular recognition as well as to the local architecture of the substrate binding site in the active site of the protease. Interactions between proteins are provided by additional protein domains in the polypeptide sequence (Figure 1.8) [43]. Configuration of these domains provides some clues to the evolutionary history of vertebrate blood coagulation.

![Protein domains of the vertebrate blood coagulation proteases.](image)

**Figure 1.8** Protein domains of the vertebrate blood coagulation proteases.
Through gene sequence analysis, gene duplication and divergence of the coagulation proteases probably occurred prior to the emergence of the vertebrate lineage. Comparison of the published gene sequences from a number of organisms ranging from jawless vertebrates to humans shows that the domain organization of the coagulation machinery is highly conserved in all vertebrates [44]. Slight variations are known to exist, however, including the absence of the contact system (FXI, FXII, and kallikrein) in fish. Doolittle has proposed that the formation of the core of the pathway occurred roughly 450 million years ago [43]. In the intervening time, a significant amount of molecular evolution has taken place resulting in numerous changes to the gene sequence thereby resulting in parallel optimization of the relevant proteins for their physiological role.

1.8 Substrate Specificity of Serine Proteases

Hydrolysis of a polypeptide chain requires proper recognition, orientation and binding of the polypeptide backbone. Thus, the residues adjacent to the scissile bond have a significant impact on the rate of hydrolysis. Nearly 30 years ago, Schecter and Berger described the substrate-protease interaction and their system has been adopted in the literature (Figure 1.9)[45]. In this model the scissile peptide bond is surrounded by subsites on the protease. Substrate amino acids are termed P (for peptide) and the subsites of the protease that interact with them are called S (for subsite). Substrate residues extending towards the N-terminus of the substrate are numbered P2, P3, P4 and so forth. Conversely, substrate residues extending towards the C-terminus are labeled P2', P3', P4' and onwards. Regions in the protease are numbered according to the substrate. For example, the P1 residue is bound in the S1 pocket. Theoretically, a large amount of variation in substrate specificity can result from
the 20 possibilities of amino acid side chains, and a wide diversity of specificity is observed in the nature.

Figure 1.9 Schecter & Berger nomenclature of protease specificity.

Broadly specific proteases recognize and act at a site dictated by a single amino acid in a polypeptide chain whereas highly specific proteases recognize a short motif consisting of three to eight amino acid residues. The biological process in which the protease is involved dictates the level of specificity. Digestive enzymes found in the gut, such as trypsin and chymotrypsin, recognize and cleave polypeptides based on the presence of a single type of amino acid (Arg/Lys and Phe/Trp/Tyr, respectively). Hence, a polypeptide substrate would typically be degraded into multiple fragments for further processing and absorption. In contrast, a number of biological processes require more specific proteolysis. As mentioned previously, the vertebrate blood clotting cascade relies on the specific cleavage of each member of the pathway to function properly. A number of other biological processes require similar levels of specificity, particularly when used for signaling purposes such as hormone and chemokine activation [46]. However, a trade off exists between the level of substrate specificity and the catalytic efficiency of the enzyme.
By demonstrating a high degree of selectivity, the preferred substrate has a slow rate of association with the enzyme. In turn, this generates a decreased catalytic rate relative to non-specific enzymes. In vivo such a scenario is not often preferred and alternate mechanisms of substrate specificity are employed. As mentioned in the discussion of the blood coagulation system, additional protein domains are associated with a protease domain to aid protein-protein interactions. Thus, physiology has placed a barrier on the level of specificity that a protease might possess.

Proteases of the blood coagulation system display a marked preference for certain amino acid side chains in the P1 to P4 positions and hydrolyze them rapidly. All coagulation proteases have trypsin-like specific at the primary (P1) position and prefer to hydrolyze peptide bonds on the C-terminal side of Arg or Lys residues [47]. Extended substrate specificity exists in all coagulation proteases in the S2 to S4 binding pockets. On the basis of their specificity, both FXa and thrombin are widely used for the site-specific cleavage of recombinant proteins after Ile-Glu-Gly-Arg and Leu-Val-Pro-Arg sequences in a polypeptide chain, respectively [48]. Similar sequences to those preferred are known to be hydrolyzed in vivo by the two proteases. Kinetic analysis of these proteases has revealed that both enzymes can effectively hydrolyze other sequences of amino acids [41,49-51]. For example, FXa was initially thought to have a strict preference for a small amino acid side chains at P2 (Gly). However, large residues (Trp, Phe) at this position are preferred in vitro. Thrombin can be genetically manipulated to prefer one of its two cleavage sites, which are different in sequence and structure [52,53]. Based on these discrepancies, it seems possible to engineer at least this level of substrate specificity into a broadly specific trypsin-like protease.
1.9 Genetic Manipulation of Serine Proteases

Great strides in biotechnology have been made in the past decade that allow for the design of enzymes with desirable properties. Examples of successful modifications made through protein engineering include increased stability or activity at the extremes of temperatures or pH, resistance to oxidation, and stability in non-aqueous environments [20]. Methods to introduce these properties involve mutagenesis of a target enzyme either through structure based design or by random mutagenesis combined with some form of genetic selection [54]. A rational design strategy requires significant amounts of information, particularly three dimensional structures of the initial enzyme as well as the knowledge of the regions that would be involved. Conversely, a randomized mutagenesis procedure combined with selection or screening requires no information of the sequence, structure or mechanism and has been widely adopted for the alteration of biochemical properties of enzymes [55].

Central to all forms of protein engineering is the creation of a novel protein by adding a novel function that was not possessed by the target protein.

Creation of a highly specific protease suffers from the practical difficulties associated with producing a kinetically worse enzyme. As mentioned previously, for a protease to exhibit a high degree of substrate specificity, some compensation in catalytic efficiency must be made. A number of studies have shown that it is possible to switch substrate specificities amongst disparate members of the protease family. For example, Hedstrom demonstrated the conversion of a trypsin-like enzyme, which prefers P1 Arg/Lys residues, into a chymotrypsin-like enzyme that prefers Phe/Trp/Tyr at P1 [56-59]. The change in specificity required mutagenesis of three surface loops in the enzyme as well as a number of other point mutations. Importantly, the regions changed do not contact the substrate directly and the
resulting enzyme is inefficient at catalyzing the hydrolysis of peptide bonds. These observations demonstrate the inherent complexity of designing improved substrate specificity in the chymotrypsin family. Other modifications of the substrate specificity in the S1 family of proteases have also been successful (Table 1.3). A number of specificity determinants have been uncovered by mutagenesis targeted to probe additional features of the enzyme, such as zymogen processing and protease-inhibitor interactions [60-64]. The wealth of biochemical and structural information available suggests the ability to design extended substrate specificity of the S2 to S4 pockets.

<table>
<thead>
<tr>
<th>Protease</th>
<th>Engineered Property</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trypsin</td>
<td>Conversion to elastase-like primary specificity</td>
<td>[65]</td>
</tr>
<tr>
<td></td>
<td>Increased P1 specificity towards Arg side chains</td>
<td>[66]</td>
</tr>
<tr>
<td></td>
<td>Increased P1 specificity towards Lys side chains</td>
<td>[67]</td>
</tr>
<tr>
<td></td>
<td>S1' Engineering to favor basic residues</td>
<td>[68]</td>
</tr>
<tr>
<td>Kallikrein</td>
<td>Modification of S2 binding pocket</td>
<td>[69]</td>
</tr>
<tr>
<td>FIXa</td>
<td>Conversion to FXa-like extended substrate specificity</td>
<td>[70]</td>
</tr>
<tr>
<td>Thrombin</td>
<td>Alteration of P2-P4 preference creating an anticoagulant protease</td>
<td>[52,53]</td>
</tr>
</tbody>
</table>

Table 1.3 Select examples of successful protein engineering of serine proteases.

1.10 *Streptomyces griseus* Trypsin

*Streptomyces griseus* trypsin (SGT) was initially purified from Pronase – a commercial preparation of secreted proteases – and characterized on the basis of its hydrolysis of N$^\alpha$-benzoyl-L-arginine ethyl ester (BAEE) and casein and its inhibition by soybean trypsin inhibitor [71]. Further characterization of its specificity and inhibition identified SGT as a
typical broad specificity trypsin-like serine protease of the S1 family that hydrolyzes polypeptide chains on the C-terminal side of basic residues (Arg and Lys) [72]. Based on sequence alignments, SGT is more similar to bovine trypsin than to other bacterial serine proteases [73,74]. The structure of SGT was subsequently determined by x-ray crystallography and refined to 1.7 Å, and revealed a three-dimensional fold that is also more similar to mammalian serine proteases than bacterial proteases [75,76]. Although similarities exist at the sequence and structural level, SGT differs from its mammalian homologues in its reduced number of amino acid insertions in the polypeptide chain when aligned by either sequence or structure (Appendix A). Moreover, SGT contains only three disulfide bonds rather than the five or six typically observed in the S1 family of proteases. These differences suggest that SGT could be used as a model scaffold to study the substrate specificity and other properties demonstrated by mammalian proteases.

1.11 Statement of Hypothesis

If all substrate specificity determinants of coagulation factor Xa are known, then their introduction into *Streptomyces griseus* trypsin will result in a protease with similar substrate selectivity.

1.12 Objectives and Outline

Although a wealth of data is available, several questions remain about the overall mechanism of specificity of serine proteases. Amongst the poorly characterized details of serine proteases are the flexibility of the active site and its influence on substrate specificity, the role of water molecules in the active site and the possibility of designing ultra-high
specificity serine proteases that can be tailored to desired reactions. In this study, SGT is developed as a model for mammalian serine protease specificity as it has similarity in both sequence and structure, and is derived from a bacterial source that should allow for production in other bacteria and hence allow genetic modification of the protein.

Based on the structural similarities observed near and around the active site of \textit{S. griseus} trypsin compared to mammalian serine proteases, site-directed mutagenesis should produce a catalytically active protease with the specificity of factor X. Four questions are to be addressed in this study:

(1) Does the recombinant SGT protein produced from \textit{Bacillus subtilis} have similar enzymatic properties as the wild type protein?

(2) What mutations are required to increase the primary specificity of the enzyme towards Arg side chains?

(3) What point mutations or surface loops near the active site confer a greater degree of extended specificity in the enzyme?

(4) What are the complete requirements to convert SGT to coagulation factor Xa-like specificity?

By engineering substrate specificity into a protease where little exists, a number of benefits will result. As mentioned previously, proteases are used in the site specific cleavage of recombinant proteins and the enzymes used are costly due to their production from blood. Proteases resulting from a bacterial expression system would cost far less and facilitate increased usage. Through the design of specificity similar to a coagulation factor, one can examine the mechanisms by which the proteases generate specificity and the roles of other regions of the polypeptide that might be involved in specificity. For example, nearly all
coagulation factor proteases have a specific sodium binding site and the binding of sodium results in a 3- to 5-fold rate increase in catalysis [77]. Sodium binding in thrombin also alters the substrate specificity of the enzyme, yet dissection of this process has been limited owing to the interconnected relationship amongst biochemical events within the enzyme.

Development of a recombinant bacterial expression system for *S. griseus* trypsin (Chapter 2) was a crucial obstacle to overcome in this research. Using this expression system, the primary specificity of the protease was genetically engineered to favor Arg over Lys side chains at the P1 position (Chapter 3). Subsequent mutagenesis of the S2 to S4 pockets of SGT was carried out to engineer coagulation factor Xa substrate specificity to the enzyme (Chapter 4). On the basis of these results, a framework for the production of novel proteases with substrate specificities not observed in nature is outlined and other future directions are discussed (Chapter 5).
2.1 Introduction

Development of an efficient, cost effective and scaleable recombinant protein expression system is the first step in protein engineering. Soluble, active and high purity protein must result from an efficient expression system. A number of organisms have been used for this purpose including those from bacteria, fungi, yeast, and eukaryotic cell lines. Protein expression in lower organisms, such as gram-positive and gram-negative bacteria, costs significantly less but the drawback of these systems is their inability to produce complex proteins. Longer production times and higher costs are associated with eukaryotic based systems; however, their use is usually required when the protein to be produced is large (>60 kDa), has a complex fold and contains disulphide bonds or requires post-translational modification (such as glycosylation). For protein engineering, bacterial, fungal, or yeast expression hosts are typically used due to their amenability to genetic manipulation and efficiency of protein production.

Successful expression of a recombinant protein in lower organisms requires a number of favorable biochemical features. Low toxicity, simple structural fold, lack of disulphide bonds, lack of post-translational modification, and small size tend to help production. Recombinant expression is also significantly influenced by the properties of the gene that encodes the polypeptide. Optimal codon usage and lack of secondary structure have been shown to be problematic in the expression of several proteins [78]. In many instances,
however, having favorable characteristics at the genetic and protein level may still not result in successful expression and a great deal of trial and error in different systems is needed.

Bacterial protein expression is widely performed in *Escherichia coli*. As a host for genetic manipulation, *E. coli* is unrivalled in the diversity and simplicity of methods established for genetic manipulation. Purification and alteration of DNA in these gram-negative bacteria is straightforward and highly reproducible. Indeed, much of the history of molecular biology is the result of the study of this organism. Unfortunately, *E. coli* is not as adept at the production of foreign proteins. A number of attempts have been made to engineer the genome of this organism to increase its capacity for recombinant protein production, yet no universal solution has been found [79]. For this reason a number of other bacteria have been investigated as alternatives including *Bacilli* [80,81], *Lactobacilli* [82], *Streptomyces* [83], *Pseudomonads* [84] and *Caulobacter* [85]. Importantly, these alternatives have the ability to secrete heterologous proteins outside of the cell. Secretion eases subsequent purification, reduces the toxicity associated with the protein, and improves the rate of formation of disulphide bonds.

On the basis of its bacterial source, *Streptomyces griseus* trypsin is suggested as a good target for recombinant protein expression and subsequent protein engineering. SGT has many features that are required for successful protein engineering. The structure of the enzyme has been determined at high resolution and a wealth of structure-function information has been described for highly similar enzymes [75,76,86]. Although it has a number of desirable properties for recombinant expression such as its small size and simple fold, several features of the gene and protein could complicate the production of the recombinant protein. Translation of the mRNA encoding the protein may be hampered by the high guanine and
cytosine content of the SGT gene (70%). Several authors have shown that sub-optimal codon usage in the first several codons can dramatically decrease protein expression [87,88]. In the reducing environment of a bacterial cell, the nascent polypeptide chain may have difficulty forming the three disulphide bonds required to stabilize the structure of SGT [89]. Lastly, if the protease is produced in an active form, it may degrade components of the cell as well other SGT polypeptides. These properties indicate that production of the recombinant protein may be difficult.

Once produced by an expression host, a recombinant protein must be purified to homogeneity. In this process, some form of protein capture to concentrate and crudely purify the protein is typically linked with one or more chromatographic separations. Throughout this procedure the loss of protein, whether due to instability or the activity of contaminants, must be minimized. Given the large body of literature on serine proteases, particularly trypsin-like proteases of the chymotrypsin family, a number of reagents are available commercially and can be used in a variety of purification methods. One particular advantage of using a bacterial expression host is the lack of post-translational modification in the target protein which minimizes sample heterogeneity in the protein purified. Thus, recombinant trypsin-like proteases derived from a bacterial source should be easily purified in high yield.

To begin introducing substrate specificity into SGT, an efficient expression system was required. I chose a bacterial expression system that would facilitate downstream processing and future high-throughput studies. *B. subtilis* is an excellent expression system for SGT due to its ability to secrete active recombinant protein into the extracellular environment. Comparison of the enzymatic properties and three dimensional structure of the purified protein to the natively derived protein from *Streptomyces griseus* show that both
proteins are identical. These studies provide the basis for the subsequent engineering of substrate specificity.

2.2 Materials & Methods

2.2.1 Plasmids, Bacterial Strains, and Growth Conditions

*Escherichia coli* was grown using standard methods [90]. *B. subtilis* strain WB700 was grown in super-rich medium [91] or on tryptose blood agar base (Difco) at 37°C. For the *B. subtilis* carrying plasmid pWB980 [92], kanamycin was added to a final concentration of 10 μg ml⁻¹ in both liquid and solid media.

2.2.2 DNA manipulation

Procedures for genomic (*S. griseus* (ATCC 10137)) and plasmid DNA manipulation were carried out using established protocols [90]. Plasmid DNA was purified using a QIAprep spin miniprep kit (Qiagen). Enzymes were obtained from New England Biolabs and Roche Molecular Biochemicals. For PCR amplification of the *SprT* gene encoding SGT, the following oligonucleotides were designed to maintain reading frame of the sacB signal peptide present in plasmid pWB980:

5'-ggaagctttgttagGGGCGGGCGGCCGAGG-3'
5'-ggtctagatttCAGCGGGGCGGCCGAGG-3'

(restriction sites are underlined, the *SprT* gene specific sequences are given in upper case). The PCR fragment was first cloned into pBluescript KS+ (Stratagene) and then sub-cloned into pWB980 using the HindIII and XbaI restriction enzyme sites contained in the oligonucleotide primers. Transformation of *B. subtilis* strain WB700 was performed by the
method of Spizizen [93]. DNA sequence analysis of the cloned gene was performed using the
Big Dye Terminator kit and analyzed on an ABI 3700 DNA Sequencer (Applied Biosystems).

2.2.3 Protein Purification

In order to purify the native and recombinant protease to homogeneity, a purification
strategy was developed at low pH to minimize autolysis. After centrifugation to remove
cellular debris (5,000 x g, 1 hr), recombinant SGT was purified from the supernatant of 1 L of
*B. subtilis* WB700 culture. Sequential ammonium sulphate fractionation was carried out at
30% and 85% saturation. The 85% (NH₄)₂SO₄ fraction pellet was resuspended in 20 mM
sodium acetate buffer pH 4.5, dialyzed against the same buffer and applied to a column (15
cm x 1.5 cm) of SP Sepharose Fast Flow (Amersham Pharmacia). After extensive washing
with 20 mM sodium acetate buffer containing 50 mM NaCl, pH 4.5, the bound proteins were
eluted with 20 mM sodium acetate buffer, pH 4.5, containing 150 mM NaCl. The active
fractions were pooled and applied to a Benzamidine Sepharose 4 Fast Flow column (8 cm x
0.75 cm) (Amersham Pharmacia). The column was washed with 20 mM sodium acetate buffer
containing 500 mM NaCl, pH 4.5, and the enzyme was eluted in the same buffer containing in
addition 40 mM benzamidine HCl (Sigma). Fractions containing active protease were pooled,
concentrated and dialyzed against 10 mM Tris-HCl buffer containing 150 mM NaCl and 20
mM CaCl₂, pH 7.6 using a 10,000 NMWL Ultrafree-4 centrifugal filter unit (Millipore). Gel
filtration through a column (45 cm x 0.75 cm) of Sephadex G-75 (Amersham Pharmacia) was
performed using 10 mM Tris-HCl containing 150 mM NaCl and 20 mM CaCl₂, pH 7.6.
Similarly, native SGT was isolated from 1 g of extracellular filtrate from *S. griseus* (Sigma).
The final protein concentration was determined by UV absorbance at 280 nm, using the
extinction coefficient 37,100 M\(^{-1}\) cm\(^{-1}\) [94] or by a BCA protein assay kit (Pierce). Active site titration was performed using 4-nitrophenyl p'-guanidinobenzoate and a standard curve of p-nitrophenol (Sigma). Sodium dodecylsulphate polyacrylamide gel electrophoresis and Coomassie Blue staining were performed according to standard procedures [90]. N-terminal protein microsequence analysis was performed by the University of Victoria - Genome BC Proteomics Centre (Victoria, Canada). Electrospray-mass spectrometry was carried out on a PE-Sciex API 300 triple quadrupole mass spectrometer (Sciex) equipped with an Ionspray ion source. The mass spectrometry was performed by Dr. S. He in the Withers laboratory (Dept. of Chemistry, UBC).

2.2.4 Kinetic Analysis

Kinetic analysis was performed in 10 mM Tris-HCl buffer containing 150 mM NaCl, 20 mM CaCl\(_2\), and 0.1 % PEG 8000, pH 7.6. A standard of 7-amino 4-methylcoumarin (AMC) was used to quantify the rates of hydrolysis of the fluorogenic substrates Tos-Gly-Pro-Arg-AMC and Tos-Gly-Pro-Lys-AMC (Bachem). A minimum of six substrate concentrations ranging from 1 to 50 \(\mu\)M was used. The final concentration of the enzyme in each assay was 0.5 nM. Non-linear regression of the initial reaction rates and calculation of the kinetic parameters were performed using the Graphpad Prism 3.0 software (Graphpad).

2.2.5 Crystallization

In previous studies, crystals of the native SGT were obtained through batch crystallization using (NH\(_4\))\(_2\)SO\(_4\) [75,76]. In the current study, proteins were crystallized using similar conditions (10-15 mg/mL protein, 1.5 M (NH\(_4\))\(_2\)SO\(_4\), 10 mM calcium acetate, pH 6.2)
except hanging drop vapor diffusion was utilized where the reservoir contained 1.55 M (NH₄)₂SO₄. Crystals appeared in two to three weeks to dimensions of approximately 0.3 x 0.3 x 0.3 mm. Data were collected at 100 K (Oxford Cryostream) with a Mar345 detector mounted on a Rigaku RU-200 X-ray generator (50 kV, 100 mA) with Osmic focusing mirrors. Crystals were soaked briefly in 20% glycerol, 2.2 M (NH₄)₂SO₄ for cryoprotection prior to data collection. Data were processed using the HKL package and refined using CNS version 1.1 in combination with Xtalview [95-97]. The previously reported native SGT structure (PDB entry 1SGT) was used as a model for rigid body refinement of the structure [76].

2.3 Results and Discussion

2.3.1 Production and Purification of Recombinant SGT

Previous studies demonstrated the ability to produce soluble trypsin-like enzymes in the periplasmic space of E. coli [98-100]. In our experiments, however, E. coli was incapable of generating soluble SGT despite using a variety of plasmid constructs in a number of bacterial host strains. The presence of three disulfide bonds, the high G+C% content of the SprT gene (70%) and the toxicity of the recombinant protein are possible reasons for the lack of production of recombinant proteins in E. coli [78,88]. To overcome these limitations B. subtilis WB700, a strain that is deficient in seven proteases, was used to produce sufficient yields of recombinant SGT for kinetic and structural analysis [101]. Unlike E. coli, B. subtilis is capable of secreting proteins into the extracellular environment, which facilitates rapid detection, purification and analysis of recombinant proteins.
Secretion of proteins into the extracellular medium is facilitated by the presence of a single plasma membrane in the gram-positive bacterium *B. subtilis*. Following translation by the ribosome, a nascent polypeptide chain is targeted for secretion by the presence of a cleavable amino-terminal signal peptide. On the basis of genome sequence analysis, over 300 proteins (~7.3% of all genes) have been postulated for secretion in *B. subtilis* [102-104]. The ability to secrete such a wide diversity of proteins by this organism has been used for the production of a number of recombinant proteins [105-107].

Protein secretion in all bacteria, including *B. subtilis*, is primarily due to the ATP-dependent Sec pathway. Recognition of the signal peptide is mediated by the signal recognition particle protein complex, which shuttles the unfolded protein to the cell membrane [108]. Removal of the signal peptide occurs as the denatured protein translocates across the cell membrane and is typically carried out by the type I signal peptidase SipS in *B. subtilis* [109]. In our expression system, secretion of SGT into the extracellular environment was mediated by fusing the gene to the signal peptide sequence of levansucrase (SacB) (Figure 2.1). Thus, the N-terminus of the protein is not accessible to the active site and the protease is inactive until it is secreted from the cell. Cleavage after the Ala-Phe-Ala sequence at the junction of the fusion protein by SipS generates the correct N-terminus. SGT can then fold in the comparatively non-reducing environment outside of the cell. By mimicking the native organism for the production of the recombinant SGT, we have developed an efficient and novel expression system for the production of trypsin-like enzymes.
Figure 2.1 Plasmid map of SGT gene cloned into pWB980 for recombinant protein expression in *B. subtilis*. The gene was cloned to maintain the reading frame of the SacB signal peptide which facilitates secretion of the recombinant protein into the extracellular environment.

Recombinant protein yields of >15 mg/L of culture medium were obtained within 24 hours of growth at 37°C. The four step purification typically produced 10-15 mg/L of *B. subtilis* culture with an overall yield of 80% (Table 2.1). Four separation techniques were applied to yield the highest purity enzyme possible. The methods were chosen based on their compatibility and mild conditions. For example, an affinity chromatography step using soybean trypsin inhibitor was found to bind SGT effectively. However, removal of the protein
from this type of column required pH 2.0 and it was feared that such conditions may destroy
unstable mutants of SGT.

<table>
<thead>
<tr>
<th></th>
<th>Vol. (mL)</th>
<th>Total Protein (mg)</th>
<th>Total Activity (mmol/s)</th>
<th>Specific Activity (µmol/s/mg)</th>
<th>Purification (fold)</th>
<th>Yield (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Media</td>
<td>1000</td>
<td>16000</td>
<td>59</td>
<td>4</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>(NH₄)₂SO₄</td>
<td>50</td>
<td>400</td>
<td>53</td>
<td>134</td>
<td>34</td>
<td>90</td>
</tr>
<tr>
<td>precipitate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SP Sepharose</td>
<td>15</td>
<td>18.3</td>
<td>51</td>
<td>2774</td>
<td>694</td>
<td>87</td>
</tr>
<tr>
<td>Benzamidine</td>
<td>5</td>
<td>12.4</td>
<td>49</td>
<td>3856</td>
<td>964</td>
<td>81</td>
</tr>
<tr>
<td>Sepharose G-75 Superdex</td>
<td>4</td>
<td>12.2</td>
<td>48</td>
<td>3955</td>
<td>989</td>
<td>81</td>
</tr>
</tbody>
</table>

**Table 2.1** Purification table for recombinant SGT (bSGT) from *B. subtilis* extracellular supernatant. Activity was measured by the hydrolysis of the chromogenic substrate Bz-Ile-Glu-Arg-pNA at 40 µM in 10 mM Tris-HCl, 20 mM CaCl₂, pH 7.6.

Native and recombinant SGT were purified to homogeneity prior to analysis. Purity was assessed by several different criteria. By electrospray ionization mass spectrometry, the expected and observed masses for the native and recombinant protein were within experimental error (23106.9 and 23107.0 amu, respectively). For both proteins the integrated data from the $m/z^+$ fragments yielded a single unambiguous peak with minimal background (Figure 2.2). In addition, when the protease was treated with phenylmethanesulfonyl fluoride, SDS-PAGE showed a single band of the expected molecular weight (Figure 2.3). Only two autolytic fragments of SGT were observed when the protease was boiled prior to SDS-PAGE in the absence of a strong inhibitor.
Amino-terminal sequence analysis revealed that the wild-type enzyme had a single unambiguous sequence NH$_2$-Val-Val-Gly-Gly-Thr-Arg corresponding to the published SGT sequence (12). Together with the protein assays and active site titration data, these results suggest that the final protein preparation was greater than 99% pure.

**Figure 2.2.** ES-MS spectrum of purified recombinant SGT. The single $m/z^+$ peak at 23,107.0 amu with minimal background indicates the high purity of the recombinant protein.
Figure 2.3 SDS-PAGE of purified recombinant SGT. Lane A: bSGT (0.5 μg) inactivated with PMSF prior to addition of SDS-PAGE loading buffer and boiling. Lane B: bSGT (1.0 μg) without PMSF inhibition. The central lane of the gel contains low-range protein molecular mass standards (Bio-Rad) whose masses are given on the left-hand side of the gel. The gel was stained with Coomassie Brilliant Blue.

2.3.2 Kinetic Analysis

Two fluorogenic peptide substrates, Tos-Gly-Pro-Arg-AMC and Tos-Gly-Pro-Lys-AMC, were used to monitor the P1 Arg:Lys preference of the native and recombinant proteases (Table 2.2). The kinetic parameters of the native and recombinant wild-type SGT are similar to previously reported values using the same pair of substrates [66]. On the basis of the similar rates of hydrolysis of these peptide substrates, we can conclude that the recombinant protein behaves identically to the native protease.
### Table 2.2

<table>
<thead>
<tr>
<th></th>
<th>Tos-Gly-Pro-Arg-AMC</th>
<th>Tos-Gly-Pro-Lys-AMC</th>
<th>S&lt;sub&gt;R/SK&lt;/sub&gt;&lt;sup&gt;†&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>k&lt;sub&gt;cat&lt;/sub&gt;</td>
<td>K&lt;sub&gt;m&lt;/sub&gt;</td>
<td>k&lt;sub&gt;cat&lt;/sub&gt; / K&lt;sub&gt;m&lt;/sub&gt;</td>
</tr>
<tr>
<td></td>
<td>(min&lt;sup&gt;−1&lt;/sup&gt;)</td>
<td>(μM)</td>
<td>(min&lt;sup&gt;−1&lt;/sup&gt; μM&lt;sup&gt;−1&lt;/sup&gt;)</td>
</tr>
<tr>
<td>SGT</td>
<td>4880 ± 410</td>
<td>2.3 ± 0.2</td>
<td>2122 ± 120</td>
</tr>
<tr>
<td>bSGT</td>
<td>4570 ± 1210</td>
<td>2.0 ± 0.2</td>
<td>2285 ± 60</td>
</tr>
</tbody>
</table>

<sup>†</sup> S<sub>R/SK</sub> = (Tos-Gly-Pro-Arg-AMC k<sub>cat</sub> / K<sub>m</sub>)/(Tos-Gly-Pro-Lys-AMC k<sub>cat</sub> / K<sub>m</sub>)

Table 2.2  P1 arginine to lysine preference of SGT enzymes. Arg:Lys preference was measured by amidolytic activity of the native (SGT), recombinant SGT (bSGT) using two fluorogenic peptides, Tos-Gly-Pro-Arg-AMC and Tos-Gly-Pro-Lys-AMC. Values obtained in triplicate ± S.D.

#### 2.3.3 Crystallization and Structure Determination

X-ray diffraction data were obtained for the recombinant wild-type SGT at 1.5 Å resolution. Data collection and refinement statistics are given in Table 2.3. The recombinant protease crystallized in the C222<sub>1</sub> space group and contained one molecule per asymmetric unit and a Matthews coefficient of 2.2 Å<sup>3</sup>/Da (Table 2.3). The structure was deposited in the PDB database as 1OSS. During refinement, low R<sub>cryst</sub> and R<sub>free</sub> values were obtained (0.19 and 0.22). These were accompanied by excellent stereochemistry indicating a high quality model. Inspection of the Ramachandran plot revealed that all non-glycine backbone atoms are in allowed regions, with only Asn178 adopting a conformation in the generously allowed region (Figure 2.4). The overall B-factors for the polypeptide atoms were low (~13 Å<sup>2</sup>), and regions with high B-factors were limited to solvent exposed regions that are not involved in crystal packing. The structure of SGT is the highest quality reported to date likely due to the
decreased radiation damage as the present crystals were analyzed under cryogenic conditions with a shorter collection time.

**Figure 2.4** Ramachandran plot of the crystal structure of recombinant wild-type SGT. All non-glycine residues are in the allowed conformation. The plot was calculated using PROCHECK [110]
Data collection

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Resolution (Å)</td>
<td>1.5 (1.55 - 1.65)</td>
<td></td>
</tr>
<tr>
<td>Total Observations</td>
<td>25923</td>
<td></td>
</tr>
<tr>
<td>Completeness (%)</td>
<td>84.1 (74.1)</td>
<td></td>
</tr>
<tr>
<td>Average redundancy</td>
<td>2.8</td>
<td></td>
</tr>
<tr>
<td>I/σI</td>
<td>20.4 (5.1)</td>
<td></td>
</tr>
<tr>
<td>R_{merge} (%)</td>
<td>4.1 (18.2)</td>
<td></td>
</tr>
</tbody>
</table>

Refinement statistics

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Space group</td>
<td>C222₁</td>
<td></td>
</tr>
<tr>
<td>Cell dimensions (a,b,c)</td>
<td>50.04, 69.82, 119.65</td>
<td></td>
</tr>
<tr>
<td>Molecules per asymmetric unit</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>R_{cryst}</td>
<td>0.196</td>
<td></td>
</tr>
<tr>
<td>R_{free}</td>
<td>0.222</td>
<td></td>
</tr>
<tr>
<td>Protein atoms†</td>
<td>1623</td>
<td></td>
</tr>
<tr>
<td>Solvent atoms per asymmetric unit</td>
<td>211</td>
<td></td>
</tr>
<tr>
<td>Average B-factor for protein (Å²)</td>
<td>12.9</td>
<td></td>
</tr>
<tr>
<td>Average B-factor for water (Å²)</td>
<td>21.0</td>
<td></td>
</tr>
<tr>
<td>Occupancy of Ca²⁺ (B Å²)</td>
<td>0.53 (14.98)</td>
<td></td>
</tr>
<tr>
<td>Bond length deviations (Å)</td>
<td>0.007</td>
<td></td>
</tr>
<tr>
<td>Bond angle deviations (°)</td>
<td>1.4</td>
<td></td>
</tr>
</tbody>
</table>

† α = β = γ = 90°; † including alternate side chain conformations

**Table 2.3** Data collection and refinement statistics of wild-type recombinant SGT. Statistics for the highest resolution shell are given in parentheses.
2.3.4 Comparison of the recombinant and Pronase-derived crystal structure

All amino acid residues in the model of the recombinant SGT were clearly identified and positioned. In 1SGT [76], several residues had weak or absent side-chain density. In the present structure, most of these densities were clearly resolved, although several residues lacked density in the terminal atoms of their side-chains (Thr20, Gln75, Lys82, Thr98, Ser236, and Arg243) indicating disorder of these solvent exposed atoms. Residues 77 and 79 were modeled as Gly and Ala in 1SGT but are two Ser residues by DNA sequence analysis [74]. These Ser residues were clearly resolved in the current electron density map. The electron density of one sulphate ion was observed in the oxyanion hole of the substrate binding site and was included in the structure. The position of this sulphate is conserved in anionic salmon trypsin (PDB entry 1BIT), bovine trypsin (1TLD), and porcine pancreatic elastase (3EST) [111-113]. Alternate conformations were observed for Gln192, which either points into the solvent or forms a pair of hydrogen bonds with the backbone NH of Gly148 of an adjacent SGT molecule. In 1SGT the same residue was noted as having high mobility [76]. Differences in the Ca$^{2+}$ binding site in the recombinant crystal structure are discussed in Chapter 3. The overall differences between the native (1SGT) and wild-type recombinant enzyme are minor (Figure 2.5), with a root mean square deviation of all 892 atoms in the C$_a$ backbone of 0.27 Å. The largest deviation in backbone occurs at the 174-loop where peptide bond of Ala177a adopts a 180° rotation compared to the native structure.
Figure 2.5 Superimposition of the C\textalpha traces of native and recombinant wild-type SGT. The two structures (PDB ID 1SGT and 1OS8) are indistinguishable as indicated by the small r.m.s deviation between the C\textalpha backbone atoms of the two structures (0.27 Å).

2.3.5 Wild-type Native and Recombinant SGT Substrate Binding

A negatively charged residue (D189) is present in the S1 pocket and confers the primary specificity of SGT and other trypsin-like enzymes towards positively charged Arg or Lys side chains [114]. Based on the structure of 1SGT, D189 is located at the base of a narrow cylindrical cleft that can accommodate these side chains [76]. The slight preference for Arg over Lys in this pocket is due to the requirement for a bridging water molecule between the shorter lysyl-side chain and D189 [66]. The γ-OH of residue T190 interacts directly with the substrate via hydrogen bonding. Both Lys and Arg side chains adopt
favorable conformations for interaction with D189 and the hydroxyl group of T190. Kinetic analysis of the native and recombinant SGT proteases demonstrated an Arg:Lys preference of 4:1.

Extended substrate specificity is largely absent in all broadly specific trypsin-like enzymes. Accessibility of the catalytic triad in SGT is not hindered by the structure of the active site. Crystal structures of other trypsin-like enzymes have demonstrated that the peptide backbone of substrate residues P1 to P3 forms an anti-parallel β-sheet with residues 214 to 216 in the enzyme [24]. Analysis of the substrate specificity of the S2 to S4 pockets in bovine trypsin has shown an absence for preference of any side chain at these positions [115,116]. The structure of SGT suggests an identical mode of substrate binding and an absence of substrate specificity in the S2 to S4 pockets.

2.4 Conclusions

A novel expression system for the production of recombinant SGT has been developed using *B. subtilis*. High purity protease resulted from a four-step purification protocol. The recombinant protein demonstrates identical biochemical and structural properties to the *Streptomyces* derived protease. On the basis of the high level of production, purity, and ability to crystallize the protein, the recombinant protein is highly amenable for engineering extended substrate specificity into the enzyme.
Chapter 3. Engineering the Primary Substrate Specificity of

*Streptomyces griseus* Trypsin

3.1 Introduction

Substrate specificity is a key concept in the analysis of serine proteases. Sequence analysis studies show that the S1 family of trypsin-like enzymes likely evolved from a common ancestral gene [117]. The culmination of incremental evolutionary steps led to the appearance of a number of highly specific proteases such as those found in the vertebrate blood coagulation cascades. These proteases fulfill regulatory roles in cellular processes that are distinct from their more primitive roles as degradative and protective enzymes [118]. Much data have been collected on the specificity determinants of serine proteases, resulting in a classification system based on their primary specificity (S1 pocket). The specificity of trypsin-like enzymes at the S1 pocket is largely defined by the presence of a negatively charged side chain at position 189 [114]. Optimal binding of the positively charged substrate (Arg or Lys side chains) to residue 189 is mediated by residue 190 [66,119]. In trypsin-like serine proteases, position 190 is occupied by a limited number of amino acids. Degradative proteases with low primary specificity (Arg:Lys preference of 4:1) display Gln, Thr or Ser at position 190, whereas proteases with high primary specificity (Arg:Lys preferences greater than 7:1) contain Ala or Ser at this position [120]. Previous studies have shown that mutagenesis of position 190 can be used to manipulate the substrate specificity of trypsin to favor cleavage after either Arg or Lys side chains [66,67,119].

Similar to degradative vertebrate trypsin-like enzymes, SGT demonstrates a primary substrate preference of Arg:Lys of 4:1. In the previous chapter I described the production of
fully active recombinant SGT from *B. subtilis*. Using this system, mutants of SGT were constructed with altered preference for arginine to lysine (Arg:Lys). Mutations were designed to mimic those found in other trypsin-like enzymes.

SGT mutant T190P is considerably more active and less Arg-specific when compared with the previously published S190P mutation created in rat anionic trypsin [67]. Kinetic and structural analysis of the mutant protease shows that both the activity and specificity of the enzyme is affected by residues surrounding residue 190. These results further our understanding of the primary substrate specificity of trypsin-like enzymes. Based on the ease of production and purification of the recombinant protein in our *B. subtilis* expression system, SGT is an ideal scaffold for the introduction of additional mutations to enhance the substrate specificity of the S2 to S4 binding pockets.

### 3.2 Materials & Methods

#### 3.2.1 DNA Manipulation and Protein Purification

Using the previously described SGT gene cloned into pBluescript KS+ plasmid, mutagenesis was performed on the gene using a QuikChange site-directed mutagenesis kit (Stratagene) as described by the manufacturer. Oligonucleotides used for mutagenesis are provided in Table 3.1. DNA sequence analysis of the cloned gene and mutants was performed using the Big Dye Terminator kit and analyzed on an ABI 3700 DNA Sequencer (Applied Biosystems). Mutant SGT genes were sub-cloned into plasmid pWB980 and transformed into *B. subtilis* WB700 as described previously. Mutant SGT proteins were expressed and purified in an identical manner as the wild-type.
<table>
<thead>
<tr>
<th>Mutation</th>
<th>Oligonucleotide</th>
</tr>
</thead>
<tbody>
<tr>
<td>T190A</td>
<td>5'-GGCGTCGACGCCTGCCAGGGGT-3'</td>
</tr>
<tr>
<td>T190P</td>
<td>5'-GGCGTCGACCCCTGCCAGGGGT-3'</td>
</tr>
<tr>
<td>T190V</td>
<td>5'-GGCGTCGACTTCTGCCAGGGGT-3'</td>
</tr>
<tr>
<td>T190V</td>
<td>5'-GGCGTCGACGTCTGCCAGGGGT-3'</td>
</tr>
</tbody>
</table>

Table 3.1 Oligonucleotides used to mutate residue 190 in the SGT gene. The reverse complement sequences of these oligonucleotides were also used in the mutagenesis.

3.2.2 Kinetic Analysis

Kinetic analysis was performed in 10 mM Tris HCl buffer containing 150 mM NaCl, 20 mM CaCl₂, and 0.1 % PEG 8000, pH 7.6. A standard of 7-amino 4-methylcoumarin (AMC) was used to quantify the rates of hydrolysis of the fluorogenic substrates Tos-Gly-Pro-Arg-AMC and Tos-Gly-Pro-Lys-AMC (Bachem). A minimum of six substrate concentrations ranging from 1 to 500 μM was used. Enzyme concentrations ranged from 0.5 to 10 nM. Benzamidine concentrations ranged from 5 to 200 μM. Non-linear regression of the initial reaction rates and calculation of the kinetic parameters were performed using the Graphpad Prism 3.0 software (Graphpad).

3.2.3 Crystallization and Structure Determination

Crystallization conditions for the T190P mutant of SGT were similar to the wild-type protein; however 25 mM benzamidine was included in the sample buffer. Crystals appeared in two to three weeks to dimensions of 0.3 x 0.3 x 0.3 mm. Data collection and structure refinement were identical to those used for the wild-type recombinant structure. Electron
density of the inhibitor in the S1 pocket of the mutant enzyme was evident, and was included in the final model.

3.3 Results and Discussion

3.3.1 Production of SGT Mutants

Expression levels of the four mutant proteases were comparable to the wild-type SGT protein suggesting that the mutations were not detrimental to the overall stability and folding of the proteases. To ensure accurate kinetics, the complete method using the four-step purification protocol was applied to each protein. ES-MS analysis of the recombinant proteases demonstrated the presence of the mutation, as well as high purity of the sample (Table 3.2).

<table>
<thead>
<tr>
<th></th>
<th>Theoretical Mass (amu)</th>
<th>Observed Mass (amu)</th>
<th>Difference (amu)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T190A</td>
<td>23077.0</td>
<td>23079.0</td>
<td>2.0</td>
</tr>
<tr>
<td>T190P</td>
<td>23103.0</td>
<td>23098.8</td>
<td>4.2</td>
</tr>
<tr>
<td>T190S</td>
<td>23093.0</td>
<td>23090.0</td>
<td>3.0</td>
</tr>
<tr>
<td>T190V</td>
<td>23107.1</td>
<td>23102.0</td>
<td>5.1</td>
</tr>
</tbody>
</table>

Table 3.2 ES-MS analysis of the four mutants of SGT.

3.3.2 Kinetic Analysis

Two fluorogenic peptide substrates, Tos-Gly-Pro-Arg-AMC and Tos-Gly-Pro-Lys-AMC, were used to monitor the P1 Arg:Lys preference of recombinant and mutant proteases (Table 3.3). Values for the native and wild-type recombinant SGT are provided in the table.
for reference. Mutants T190P and T190A demonstrated a significant increase in P1 Arg:Lys preference over the wild-type enzyme of 18:1 and 8:1, respectively. All four mutants showed increased $K_m$ values for the substrates tested suggesting that the S1 pocket of SGT is optimized for substrate binding, a feature that has been observed in other trypsin-like enzymes [66,67,119]. A similar trend of fold differences for the $K_i$ value of the small molecule inhibitor benzamidine was observed, with the exception of the T190P mutant (Table 3.4).

<table>
<thead>
<tr>
<th></th>
<th>Tos-Gly-Pro-Arg-AMC</th>
<th>Tos-Gly-Pro-Lys-AMC</th>
<th>$\frac{S_R}{S_K}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$k_{cat}$ (min$^{-1}$)</td>
<td>$K_m$ (μM)</td>
<td>$k_{cat} / K_m$ (min$^{-1}$ μM$^{-1}$)</td>
</tr>
<tr>
<td>SGT</td>
<td>4880 ± 410</td>
<td>2.3 ± 0.2</td>
<td>2122</td>
</tr>
<tr>
<td>bSGT</td>
<td>4570 ± 1210</td>
<td>2.0 ± 0.2</td>
<td>2285</td>
</tr>
<tr>
<td>T190A</td>
<td>4950 ± 470</td>
<td>12.9 ± 1.0</td>
<td>384</td>
</tr>
<tr>
<td>T190P</td>
<td>3610 ± 130</td>
<td>67 ± 5</td>
<td>54</td>
</tr>
<tr>
<td>T190S</td>
<td>6036 ± 561</td>
<td>6.1 ± 0.6</td>
<td>990</td>
</tr>
<tr>
<td>T190V</td>
<td>2300 ± 177</td>
<td>224 ± 20</td>
<td>11</td>
</tr>
</tbody>
</table>

$S_R/S_K = \frac{(\text{Tos-Gly-Pro-Arg-AMC } k_{cat} / K_m)}{(\text{Tos-Gly-Pro-Lys-AMC } k_{cat} / K_m)}$

Table 3.3 P1 Arginine to Lysine preference of mutant SGT enzymes. Arg:Lys preference was measured by amidolytic activity of the native (SGT), recombinant (bSGT) and mutants of SGT using two fluorogenic peptides, Tos-Gly-Pro-Arg-AMC and Tos-Gly-Pro-Lys-AMC. Values obtained in triplicate ± S.D.
Table 3.4 $K_i$ values of benzamidine for recombinant SGT and the four mutant forms.

<table>
<thead>
<tr>
<th></th>
<th>$K_i$ (μM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>bSGT</td>
<td>2.7 ± 0.8</td>
</tr>
<tr>
<td>T190A</td>
<td>9.9 ± 0.7</td>
</tr>
<tr>
<td>T190P</td>
<td>16.4 ± 2.2</td>
</tr>
<tr>
<td>T190S</td>
<td>4.8 ± 0.8</td>
</tr>
<tr>
<td>T190V</td>
<td>197 ± 37</td>
</tr>
</tbody>
</table>

3.3.3. Crystallization and Structure Refinement

X-ray diffraction data were obtained for the T190P mutant of SGT at 1.9 Å resolution. Data collection and refinement statistics are given in Table 3.5. As with the recombinant wild-type, the mutant crystallized in the C222₁ space group, contained one molecule per asymmetric unit and a Matthews coefficient of 2.3 Å³/Da. The mutant structure was deposited in the PDB database as 1OS8. During refinement, low $R_{cryst}$ and $R_{free}$ values were obtained (0.17 and 0.21). These were accompanied by excellent stereochemistry (Figure 3.1).
<table>
<thead>
<tr>
<th>Data collection</th>
<th>T190P SGT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resolution (Å)</td>
<td>1.9 (1.93 - 2.05)</td>
</tr>
<tr>
<td>Total Observations</td>
<td>14471</td>
</tr>
<tr>
<td>Completeness (%)</td>
<td>89.7 (79.1)</td>
</tr>
<tr>
<td>Average redundancy</td>
<td>7.1</td>
</tr>
<tr>
<td>I/σI</td>
<td>44.4 (25.2)</td>
</tr>
<tr>
<td>Rmerge (%)</td>
<td>3.1 (5.6)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Refinement statistics</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Space group</td>
<td>C222₁</td>
</tr>
<tr>
<td>Cell dimensions (a,b,c)</td>
<td>50.08, 69.60, 119.83</td>
</tr>
<tr>
<td>Molecules per asymmetric unit</td>
<td>1</td>
</tr>
<tr>
<td>R_{cryst}</td>
<td>0.167</td>
</tr>
<tr>
<td>R_{free}</td>
<td>0.212</td>
</tr>
<tr>
<td>Protein atoms</td>
<td>1632</td>
</tr>
<tr>
<td>Solvent atoms per asymmetric unit</td>
<td>193</td>
</tr>
<tr>
<td>Average B-factor for protein (Å²)</td>
<td>12.8</td>
</tr>
<tr>
<td>Average B-factor for water (Å²)</td>
<td>21.4</td>
</tr>
<tr>
<td>Occupancy of Ca^{2+} (B Å²)</td>
<td>0.43 (17.50)</td>
</tr>
<tr>
<td>Bond length deviations (Å)</td>
<td>0.007</td>
</tr>
<tr>
<td>Bond angle deviations (°)</td>
<td>1.4</td>
</tr>
</tbody>
</table>

* α = β = γ = 90°; † including alternate side chain conformations

**Table 3.5** Data collection and refinement statistics for the T190P mutant of SGT in complex with benzamidine. Statistics for the highest resolution shell are given in parentheses.
Figure 3.1 Ramachandran plot of the crystal structure of the T190P mutant of SGT. All non-glycine residues are in the allowed conformation. The plot was calculated using PROCHECK [110]

3.3.4 Ca$^{2+}$ binding site of B. subtilis derived SGT

In the previously reported structure (1SGT) the calcium binding site consisted of the Asp165 and Glu230 carboxylate groups, two well ordered water molecules and the carbonyl oxygen atoms of residues Ala177a and Glu180 (Figure 3.2) [76]. In the present structures, Asp165 was shown to adopt a different conformation than observed previously. In the wild-type recombinant SGT structure, the carboxylate of Asp165 forms a bidentate electrostatic interaction with Arg169 rather than the structural Ca$^{2+}$ ion. In the T190P model, Asp165 adopts a pair of alternative conformations, either facing the Ca$^{2+}$ ion or towards Arg169.
Modeling of the two positions at half occupancy generated a lower B-factor for the conformation involved in the electrostatic interaction with Arg169. In both structures, the carbonyl oxygen of Ala177a was oriented away from the Ca\(^{2+}\) ion and does not appear to be involved in the interaction. Three ordered water molecules are associated with the ion, rather than the two previously observed. In contrast, a single disordered water molecule with a high B-factor (41.4 Å\(^2\)) was found near the Ca\(^{2+}\) ion in the T190P crystal structure. In other trypsin-like proteases, Ca\(^{2+}\) ions have been found previously with high B-factors relative to the overall structure suggesting less than full occupancy (25). The reduced occupancy of the ion is not surprising as a related enzyme, *Streptomyces erythraeus* trypsin, lacks Glu230 and no calcium is evident in the structure [121]. Moreover, the suggested Ca\(^{2+}\) binding site in SGT is completely different than that observed in mammalian trypsin-like enzymes [76]. When taken with previous observations that the calcium ion plays no role in catalysis but plays a role only in structural stability in low/high pH solutions, these data suggest that the calcium binding site has weak affinity for the ion [72].

### 3.3.5 T190S and the Loss of γ-CH\(_3\)

Degradative proteases involved in digestive and protective functions typically possess Ser or Thr residues at position 190. However, proteases exhibiting higher substrate specificity may also possess these residues [120]. The T190S mutant of SGT demonstrated no significant increase in overall substrate specificity, yet a minor increase in catalytic activity (\(k_{cat}\) increase of 25%) and a 3-fold increase in \(K_m\) (Table 3.3) were observed for both the Arg and Lys containing substrates. Loss of the γ-CH\(_3\) is unlikely to significantly increase the solvent accessibility of D189, and it is more likely that the increased mobility of the γ-OH results in
the increased $K_m$ values for both substrates. This is valid if $k_{cat}$ is used as a measure of the stabilization of the transition state. In eukaryotic trypsin-like enzymes that contain Ser at position 190, the space that would be occupied by a $\gamma$-CH$_3$ is filled by methyl groups from either residue 16 or 138 usually in the form of isoleucine side chains. In SGT, both of these residues are valine and possess one fewer methyl group.

![Figure 3.2](image)

**Figure 3.2** Comparison of the Ca$^{2+}$ binding site in SGT enzymes: (A) native, (B) wild-type recombinant and (C) T190P mutant of SGT. The number of water molecules (●) that co-ordinate the structural calcium ion (○), as well as the conformation of the amino acid ligands differ in all three structures.
3.3.6 T190V and the Effect of a Branched Side Chain

In contrast to the kinetics of T190S, the T190V mutant demonstrated a 2-fold reduction in $k_{\text{cat}}$ in combination with nearly a 100-fold increase in $K_m$ for both Arg and Lys containing substrates (Table 3.3). Replacement of the $\gamma$-OH with a methyl group removes the hydrogen bonding capacity of the residue and reduces the solvent accessibility of D189. This results in a destabilized transition state complex relative to the wild-type protease and a weaker electrostatic interaction between the substrate and enzyme. The 100-fold increase in $K_m$ for Arg containing substrates compared to the 70-fold increase in $K_m$ for Lys containing substrates indicates the small increase in volume has a more significant effect on the longer and bulkier arginyl side chain. D189 is not typically observed with V190 or I16 in naturally occurring tryps-in-like enzymes likely due to the poor catalytic efficiency of these combinations of side chains.

3.3.7 T190A and the Loss of $\gamma$-OH

A majority of vertebrate enzymes involved in physiological regulation, such as the coagulation factor serine proteases, possess Ala at position 190 and exhibit Arg:Lys substrate specificities ranging from 7:1 for coagulation factor Xa to greater than 14:1 for bovine thrombin [119]. The T190A mutation in SGT demonstrates the molecular basis for the predominance of Ala at this position in highly specific proteases favoring a P1 Arg residue. Optimal rates of catalysis at low concentrations of substrate tend to be requisite characteristics of these vertebrate enzymes. Kinetic analysis of the T190A mutant revealed no change in $k_{\text{cat}}$ for Arg containing substrates but rather a 10-fold reduction in $k_{\text{cat}}$ for Lys containing substrates (Table 3.3). As noted previously, the increase in solvent accessibility of D189
should stabilize the transition state complex. However, the loss of hydrogen bonding (previously provided by the γ-OH) exhibits a more pronounced effect on the Lys containing substrate due to its requirement of an ordered bridging water molecule to D189.

3.3.8 T190P

Unlike the T190A mutation, whose effects were predominantly on the \( k_{cat} \) of the Lys-containing substrate, the T190P mutation affected the \( K_m \) significantly. The \( K_m \) for the Arg containing substrate was 35-fold higher than the wild-type, compared to the 46-fold increase for the Lys containing substrate (Table 3.3). Similarly, the reduction in \( k_{cat} \) for the Arg substrate (25%) was significantly less than the Lys substrate (66%). Together, these changes generate an overall Arg to Lys preference of 18 to 1. A previous report [66] analyzed the S190P mutant of rat anionic trypsin, which is analogous to the T190P constructed in SGT. When using the same pair of substrates used in this study, wild-type rat anionic trypsin exhibits a similar primary substrate specificity to SGT. However, the S190P mutation in rat anionic trypsin results in a highly specific protease that favors the Arg substrate 135-fold over the Lys containing substrate. This increase in substrate specificity was combined with a greater than 10-fold reduction in \( k_{cat} \) for both Arg and Lys substrates. These authors suggested that Tyr228 may be involved in steric clashing with the proline ring at position 190, leading to reduced activity. In SGT, residue 228 is also Tyr suggesting an alternate binding mode of this mutant. To address these discrepancies, the crystal structure of the T190P mutant in complex with the small molecule inhibitor benzamidine was investigated.

Binding of the benzamidine inhibitor to the T190P mutant is nearly identical to that observed in other trypsin-like proteases [112]. The proline ring of residue 190 does not adopt
a conformation that occludes the negatively charged carboxylate group of Asp189, nor does it conflict with Tyr228 suggesting that the previously characterized specificity of the S190P mutant in rat anionic trypsin was the result of second shell residues at positions 16 or 138 (Figure 3.3). The mutation does not significantly affect the conformation of any of the residues surrounding T190P, including the critical Asp189. The local r.m.s. deviation is low (0.70 Å) for all 96 atoms within a 5 Å radius of residue 190. However, the backbone carbonyl group of Asp189 is rotated 45° relative to the wild-type structure due steric constraints of the proline residue. Rotation of this carbonyl group does not disrupt the hydrogen bond with the backbone nitrogen of residue 17. Hence, the moderate increase in Arg to Lys substrate specificity of this mutant is the likely the result of the strengthened interaction between the substrate and Asp189 in a more hydrophobic environment. The effect on lysyl-side chains is more predominant due to the lack of hydrogen bonding of the proline ring to the substrate or a bridging water molecule and would reduce the rate of association of the side chain with Asp189.
Figure 3.3 Comparison of the S1 binding pocket in SGT enzymes: (A) the recombinant wild-type structure and (B) the T190P mutant of SGT complexed with the benzamidine inhibitor (Benz). In the wild-type structure, the γ-OH points towards the S1 pocket and provides a H-bonding group for the substrate. Mutation of residue 190 to proline removes this H-bonding capacity without disrupting the critical D189.

Although Asp189 adopts a similar conformation to that found in the wild-type protease, it is possible that the binding of the benzamidine inhibitor stabilizes the conformation of this side chain through formation of the electrostatic interaction. Analysis of
the inhibition constants of benzamidine with the wild-type recombinant protease and four mutants reveals the basis for the increased specificity without loss of catalytic activity. Similar fold differences for the $K_m$ values for the peptide substrates and $K_i$ values relative to the wild-type are observed for all mutants except T190P. The $K_i$ value is six-fold higher than the wild-type, whereas the $K_m$ values for the Arg and Lys containing substrates increase 35-fold and 46-fold, respectively (Tables 3.3 & 3.4). As the inhibitor forms a direct electrostatic interaction with Asp189, the minor difference of the inhibitory constants suggests that in the T190P mutant the interaction occurs in a more hydrophobic environment and is not accompanied by structural rearrangement of the S1 binding pocket.

Whereas T190A demonstrates the interactions present at the S1 binding pocket in the majority of proteases with a P1 Arg preference, the kinetic analysis of T190P suggests a potential intermediate in the evolution of the vertebrate coagulation cascade. Only two natural proteases have been identified that possess Asp189 and Pro190 - hagfish prothrombin and human kallikrein 10, yet neither protein has been characterized with respect to substrate specificity. Both genes have been characterized by DNA sequence analysis of a number of overlapping cDNA library clones [122-124]. The presence of Pro at this position suggests a similar specificity as that observed for the T190P of SGT. Moreover, the residues surrounding residue 190 are identical to those found in SGT. DNA sequence analysis of the hagfish prothrombin gene reveals that many of the features attributed to substrate specificity, such as the 60- and 99-loops, are identical. Hence, the reduced catalytic activity and higher $K_m$ values caused by the proline at this position may be compensated by an increased concentration of the enzyme or substrate in the blood stream of this primitive vertebrate.
3.3.9 Second Shell Residues

A number of catalytic or structural studies involving the structure based design of enzyme properties have demonstrated an important role for second shell residues surrounding the mutation of interest [125]. In serine proteases, residue 190 interacts directly with the side chains of residues 16 (the N-terminus of the protein), 138 and 228. Residue 228 is a highly conserved tyrosine in known trypsin-like enzymes. Residues 16 and 138 are restricted to hydrophobic side chains (Val, Ile, and Leu). The differences in kinetic parameters observed between mutations made in this study, rat anionic trypsin and human trypsin (type I) are likely due to the presence or absence of methyl groups within this pair of residues [66,67,119]. Moreover, mutagenesis of residue 16 in rat anionic trypsinogen II has been demonstrated to affect primary substrate specificity of the enzyme [126]. A more detailed investigation of these residues is required to understand the basis of substrate specificity within this family of important enzymes.

3.4 Conclusions

On the basis of the ease of production of SGT in the *B. subtilis* expression system, it is possible to enhance the specificity at the S2 to S4 binding pockets by either structure-based design or a directed evolution strategy. Introduction of an affinity tag, such as a hexahistidine tag, would speed the purification process of the protein and facilitate characterization of recombinant mutant proteases. The ability to crystallize this molecule readily supports SGT as a model scaffold for understanding the mechanisms of substrate specificity in the highly evolved serine protease family.
Chapter 4. Engineering Coagulation Factor Xa Substrate

Specificity into *Streptomyces griseus* Trypsin

4.1 Introduction

4.1.1 Overview

Architecture of the active site plays a key role in the physiological functions of serine proteases. Although the catalytic machinery is similar, if not identical, within the family of serine proteases, the residues comprising the active site dictate function. Differences in the active site lead to substrate specificity and the level of regulation by protease inhibitors. Detailed understanding of the molecular basis of substrate specificity is needed to design inhibitors for therapeutic applications. Moreover, the ability to tailor protease specificity to meet specialized needs is an attainable goal. At present, it has not been established what maximum levels of substrate specificity could exist on the serine protease scaffold. Few studies have involved improving the extended substrate specificity of serine proteases.

Proteases have been extensively characterized with respect to enzymatic properties. Rates of catalysis against numerous libraries of peptide and polypeptide substrates that are both natural and synthetic in origin are readily available [41,127-131]. Inhibition constants are similarly abundant due to the importance of physiological regulation and for impeding the progression of a pathology or preventing one from developing [132-137]. In addition, a vast amount of sequence and structural data can be found in the databases. Indeed, the quantity of information compiled on proteases is daunting. Ultimately these data reflect properties that
result from approximately fifty amino acids that create and surround the active site of serine proteases.

Less than twenty amino acid residues are involved in enzyme-substrate interactions in the S1 peptidases [24]. A number of residues involved in substrate binding exert their influence via backbone contacts or by stabilization of the entire structure of the protease domain. Some examples include residues 214 to 216 which are involved in formation of the anti-parallel β-strand between enzyme and substrate and three disulfide bonds between residues 42 and 58, 168 and 182, and 191 and 220 that stabilize the entire domain. In both instances, these residues are highly conserved throughout the serine protease family and do not significantly modulate specificity, but rather assist in the enzyme-substrate interactions and stabilize the transition state of the catalytic process [138,139]. Thus, variation in substrate specificity can be reduced in complexity to roughly ten amino acid positions. These residues are located on all sides of the active site and their effects can be altered by their local environments and proximal residues.

Extended substrate specificity is a hallmark of the serine proteases of vertebrate blood coagulation. These proteases serve as useful models to understand substrate specificity throughout the entire family of S1 peptidases. Coagulation factor Xa (FXa) is important due to its central role in coagulation and wide spread use in biotechnology-related applications requiring site specific proteolysis. The predominant feature of FXa substrate specificity is a two residue loop at position 99 [140]. Insertion at this site in the sequence positions a large aromatic side chain, Tyr99, in the S2 pocket and restricts P2 substrate residues to small Gly side chains. Less is known about the other determinants of substrate specificity in FXa.
Kinetic data for the hydrolysis of a number of different peptide substrates suggests that the S3 and S4 pockets of FXa display minimal selectivity [41]. The S3 specificity pocket is typically the least stringent of the pockets throughout the S1 peptidase family due to the solvent exposure of the P3 side chain. Moderately sized hydrophobic amino acids are preferred by the S4 binding pocket which is dominated by hydrophobic side chains [116]. Given the selectivity of these pockets, it seems possible to engineer FXa-like specificity into a broadly specific protease. Moreover, it is likely that a protease with a stricter preference for the Ile-Glu-Gly-Arg FXa cleavage sequence could be engineered.

Measures of substrate specificity of a protease can differ substantially based on the type of substrate employed. Proteolytic cleavage of a short peptide can misrepresent the rate of hydrolysis of a single bond in a folded protein. Detailed characterization of FXa using comprehensive peptide substrate libraries revealed that large, planar hydrophobic residues were preferred in the S2 pocket over small residues [41,42]. Such an observation contrasts with sequence analysis of in vivo FXa substrates. Amino acid sequences of prothrombin, the predominant in vivo substrate of FXa, from organisms spanning 450 million years of vertebrate evolution show conservation of the Ile-Glu/Asp-Gly-Arg recognition sequence [122,123]. Moreover, numerous studies have employed FXa for the site specific proteolysis of recombinant proteins [127]. Reports of non-specific proteolysis by FXa are obviously not easy to find in the literature. However, the large number of successful results confirms selectivity for the cleavage sequence is high.
4.1.2 Choice of Mutations to Mimic FXa-like Specificity

Previously, I described the design of a protease with a high primary specificity for Arg in the P1 position using SGT. Using the optimal mutant of SGT from this work as a starting point, the design of FXa-like specificity was pursued. To convert the specificity of SGT into FXa, six mutations were introduced into SGT to enhance selectivity of the S2, S3 and S4 pockets (Table 4.1). Mutations were chosen based on sequence conservation in known FXa proteins from various species and inspection of various x-ray crystal structures of members of the S1 peptidase family. These mutations in SGT are involved in enzyme-substrate interactions as well as optimization of the overall architecture of the active site (Figure 4.1). Many of the mutations created also mimic what is found in activated protein C (aPC), factor IXa (FIXa), factor Xla, and factor VIIa (FVIIa).

<table>
<thead>
<tr>
<th>Name</th>
<th>Mutations</th>
</tr>
</thead>
<tbody>
<tr>
<td>T190P</td>
<td>T190P</td>
</tr>
<tr>
<td>LP</td>
<td>99-loop, T190P</td>
</tr>
<tr>
<td>YP</td>
<td>99-loop, T99Y, T190P</td>
</tr>
<tr>
<td>YFP</td>
<td>99-loop, T99Y, N174F, T190P</td>
</tr>
<tr>
<td>YSFP</td>
<td>99-loop, T99Y, Y172S, N174F, T190P</td>
</tr>
<tr>
<td>YSFMP</td>
<td>99-loop, T99Y, Y172S, N174F, E180M, T190P</td>
</tr>
</tbody>
</table>

Table 4.1 Mutants of SGT constructed to mimic the substrate specificity of FXa. For the 99-loop mutation, two residues (Lys and Glu at positions 96 and 97) were inserted similar to that found in the FXa polypeptide sequence.
Selectivity of the S2 pocket in FXa is generated by the insertion of two amino acid residues at position 99. Extension of the polypeptide chain at this location positions the Tyr99 side chain into the S2 pocket and restricts access to small aliphatic side chains [140]. Two mutations in SGT were required to engineer selectivity for P2 side chains. First, a two residue insertion was created. Second, Thr99Tyr was added to the loop construct to constrain the S2 pocket in a similar fashion to FXa.

![Figure 4.1](image)

**Figure 4.1** Residues involved in the extended substrate specificity of coagulation proteases. The Na$^+$ binding site is known to play a role in the substrate specificity of thrombin and lies adjacent to the S1 pocket.

E180M was introduced into SGT to remove the electrostatic effects of the negatively charged Glu residue and mimic a residue conserved in the entire S1 family peptidases. Sequence analysis shows that Met180 is highly conserved and exists in approximately 60% of the family. However, a functional role for residue 180 has not been described in the literature. Residue Glu180 in SGT is one of several negatively charged side chains near the S4 pocket.
that generates an overall negative electrostatic potential in the region. As hydrophobic P4 residues are preferred in FXa, it was conjectured that removal of the charged moiety in the S4 pocket would improve substrate binding.

Y172S and N174F mutations in SGT were constructed to facilitate proper positioning of the 172-loop and increase the hydrophobicity of the S4 pocket. In SGT, Tyr172 is buried in the core of the structure in a similar fashion to other trypsin-like enzymes [76]. However, Ser172 in FXa has a conformation that exposes the side chain to solvent and is involved in the positioning of residue 174 [141]. Mutation at the equivalent positions in bovine trypsin led to a flexible 172-loop and suggests replacement of residues 172 to 174 is not sufficient to create the S4 pocket of FXa [142]. Thus, stabilization of the loop should be important for optimization of the extended substrate specificity in FXa.

Y217E was created in SGT to secure the 172-loop conformation observed in FXa. Residue 217 is poorly conserved in the S1 family, yet conserved in all coagulation proteins identified from a variety of vertebrate species [122,123]. In the crystal structure of human FXa, residue 217 forms two charge assisted hydrogen bonds with Ser173 via the hydroxyl group of the side chain and amide group of the backbone [143]. Unfortunately, the Gly173Ser mutant of SGT has not been characterized yet.

Using B. subtilis as an expression host, a number of mutants at the residues discussed above were constructed in SGT and their enzymatic properties determined. Substrate specificity was investigated by comparison of the rates of hydrolysis of a small library of commercially available chromogenic peptides. The mutant bearing all seven mutations did not produce a protease with FXa-like specificity, but rather specificity towards coagulation factor XIa (FXIa) substrates was observed. An intermediate mutant with six mutations led to a
protease with moderate FXa-like specificity. These results confirm previous studies that demonstrated a role for the 99-loop and 172-loop in FXa and related enzymes. Residue 217 is confirmed as a determinant of substrate specificity. However, questions about the detailed role of residue 217 are raised by the data. Further, residue 180 has been identified as a determinant of the substrate specificity in the S1 family of peptidases.

4.2 Materials & Methods

4.2.1 Plasmids, Bacterial Strains, and Growth Conditions

_E. coli_ was grown using standard methods [90]. Plasmid DNA was purified using a QIAprep spin miniprep kit (Qiagen) and manipulated using standard protocols [90]. Enzymes were obtained from New England Biolabs and Roche Molecular Biochemicals. _B. subtilis_ strain WB700 was grown in super-rich medium [91] or on tryptose blood agar base (Difco) at 37°C. For the _B. subtilis_ carrying plasmid pWB980 [92], kanamycin was added to a final concentration of 10 μg mL⁻¹ in both liquid and solid media.

4.2.2 Construction of a Hexahistidine-tagged SGT

Plasmid SGT pET-28(a)+ was previously constructed in an unsuccessful attempt to express recombinant SGT by cloning the gene into the XbaI and XhoI restriction endonuclease cleavage sites of the plasmid. Inspection of electrospray mass spectrometry data of aged samples of recombinant SGT from _B. subtilis_ suggested that Arg243 was weakly susceptible to autolysis. The mutation Arg243Ser was constructed using the QuikChange site-directed mutagenesis kit (Stratagene) as suggested by the manufacturer using the oligonucleotide 5’- GCCTCGGCCGAGGCACGCTCGAGC-3’ and reverse complement
oligonucleotide. The C-terminal portion of the SGT gene was sub-cloned from pET28(a)+
into bSGT pWB980 via PvuII and AvrII cleavage sites (Figure 4.2). Thus, SGT was cloned in
frame with a hexa-histidine tag (HisR pWB980). As this plasmid construct was to be used for
sub-cloning mutants of SGT from the pBluescript KS+ E. coli plasmid, a small portion of the
gene was deleted by digesting with NarI and self-ligation to yield ΔNarI HisR pWB980. As a
result, successful sub-cloning of mutants would restore protease activity.

Figure 4.2 Plasmid construction for the production of recombinant His-tagged SGT
(HisR pWB980) and a deletion mutant construct for easily identifying successful sub-
cloning of mutant SGT genes (ΔNarI HisR pWB980).
4.2.3 Sequence analysis of the S1 Family peptidases

All sequences for the S1 family sub-family A peptidases were downloaded from the MEROPS database (http://www.merops.ac.uk, release date 04-03-2002) [22]. Sequences were pre-aligned by the curators of the database. The complete alignment of 740 proteases was pasted into Microsoft Excel, and then converted into single columns corresponding to individual residues in the polypeptide chain. Distribution of amino acids at each position of interest was obtained using the COUNTIF function within the program. The file served as a useful database that linked amino acids at a defined position in the polypeptide sequence to characterized enzymes. Regions where insertions or deletions occurred in the family were not handled well by this method and inspection of known crystal structures and published literature were required.

4.2.4 DNA Manipulation

Using the previously described SGT gene cloned into pBluescript KS+ plasmid (Chpt. 2.2.2), mutagenesis was performed on the gene using a QuikChange site-directed mutagenesis kit (Stratagene) as described by the manufacturer. Oligonucleotides used for mutagenesis are provided in Table 4.2. DNA sequence analysis of the cloned gene and mutants was performed using the Big Dye Terminator kit and analyzed on an ABI 3700 DNA Sequencer (Applied Biosystems). Mutant SGT genes were sub-cloned into plasmid ΔNarI HisR pWB980 via HindIII and PvuII restriction sites and transformed into B. subtilis WB700 by the method of Spizizen [93].

<table>
<thead>
<tr>
<th>Mutation</th>
<th>Oligonucleotide</th>
</tr>
</thead>
<tbody>
<tr>
<td>99-Loop</td>
<td>5'-CAGGCCCCCGGCTACAACAAGGAGGCCACCGGCAAGGACTGG-3'</td>
</tr>
<tr>
<td>T99Y</td>
<td>5'-CAAGGAGGGCTACGGCAAAGGAC-3'</td>
</tr>
<tr>
<td>E180M</td>
<td>5'-CTCGTGGCAACGAGATGATCTGCAGCGGCGATAC-3'</td>
</tr>
<tr>
<td>N174F</td>
<td>5'-TCCGGCTACGGGCTCDGAGTCGTGGCC-3'</td>
</tr>
<tr>
<td>Y172S</td>
<td>5'-GCCGCTCGCGTCCCGGCTTCGAGCT-3'</td>
</tr>
<tr>
<td>N174F</td>
<td>5'-AGCTGGGCGAGGGCTGCGCC-3'</td>
</tr>
</tbody>
</table>

Table 4.2 Oligonucleotides used to mutate the SGT gene to mimic residues found in FXa. The reverse complement sequences of these oligonucleotides were also used in the mutagenesis.

4.2.5 Purification of His-tagged SGT and Mutants thereof

Overnight *B. subtilis* 20 mL cultures were used to inoculate 250 mL of super-rich broth containing 10 μg/mL kanamycin and grown for 16 hrs at 37°C. The supernatant was harvested by centrifugation (30 min., 5000 rpm) and then passed over a Talon affinity resin column (BD Biosciences) (10 cm x 0.75 cm) equilibrated in wash buffer (50 mM sodium phosphate, 500 mM NaCl, pH 8.2). The column was washed with 10 column volumes of wash buffer and the recombinant protein eluted with wash buffer containing 100 mM imidazole. Active fractions containing recombinant protein were concentrated using a 10,000 NMWL Ultrafree-4 centrifugal filter unit (Millipore) and dialyzed with 100 mM Tris-HCl, 150 mM NaCl, 20 mM CaCl₂, pH 7.6 in the same unit. Recombinant proteins were stable at 4°C for months. Protein quantification was identical to that described previously (Chpt. 2.2.3).
4.2.6 Characterization of Substrate Specificity

Kinetic analysis was performed in 10 mM Tris-HCl buffer containing 150 mM NaCl, 20 mM CaCl₂, and 0.1 % PEG 8000, pH 7.6 at 25°C. Reactions (300 μl) were prepared in 96-well microplates using either a RoboSeq 4204 or RoboGo laboratory automation system (MWG Biotech AG, Ebersburg, Germany) and measured using a Labsystems Multiskan Ascent plate reader. Peptide substrates were handled as suggested by their manufacturers (Diapharma, American Diagnostica). Assessment of specificity using chromogenic peptide substrates was estimated by direct analysis of the hydrolysis of each substrate at 40 μM at two stages of the purification. In this method, the specificity constant (k_cat/K_m) was determined from the slope of the natural logarithm of substrate remaining as function of time. Detailed kinetic analyses were performed on a minimum of six substrate concentrations ranging from 20 to 600 μM and enzyme concentrations of 10 to 70 nM. Higher substrates concentration were not examined due to solubility difficulties with the peptides.

4.2.7 Macromolecular Substrate Specificity

Human prothrombin and FXa were purchased from Haematological Technologies. Prothrombin (3.5 μg) was digested with mutants of SGT (60 ng) and FXa (60 ng) overnight at room temperature in 10 mM Tris-HCl buffer containing 150 mM NaCl, 20 mM CaCl₂, and 0.1 % PEG 8000, pH 7.6. Proteolytic fragments were resolved by SDS-PAGE following a standard protocol [90].
4.3 Results & Discussion

4.3.1 Production of His-tagged SGT

In order to facilitate purification and analysis of a large number of mutants, simplification of the four step purification scheme described previously was required (Chpt. 2.2.3). Initial attempts to produce recombinant SGT in *E. coli* were unsuccessful, yet produced a construct bearing SGT in frame with a hexa-histidine tag in the pET-28a(+) plasmid. Several histidine residues in succession promote binding to metal ions such as Cu$^{+2}$, Ni$^{+2}$, or Co$^{+2}$. As the metal ions can be immobilized onto an appropriate chromatography medium, capture of recombinant protein from complex samples is greatly simplified. The C-terminal region of SGT bearing the tag was amenable for subcloning from pET-28a(+) into the *B. subtilis* plasmid pWB980 through a fortuitous AvrII restriction site. To ensure the tag remained on the protease, the Arg243Ser mutation was introduced into SGT to protect against potential autolytic cleavage.

Yields of SGT bearing a hexa-histidine tag at the C-terminus of the protein were three to five times lower than the non-tagged construct. Maximal yields of recombinant protein bearing the tag did not exceed 5 mg/L of culture compared to 15 mg/L for the wild-type construct. Binding of the recombinant protease to various commercially available Ni$^{+2}$ or Co$^{+2}$-chelated resins was poor suggesting that the tag was partially buried in the protein. Inspection of the crystal structure of SGT suggests that the first two His residues may lie in a shallow cleft on the enzyme surface. However, high purity protein resulted from the purification (Figure 4.3). The reduced yield of SGT bearing the tag may hinder folding of the enzyme. Alternatively, the tag may promote interactions with the negatively charged peptidoglycan found on the cell wall of the bacteria. However, the ease of protein purification...
based on the tag outweighed the demand for greater amounts of protein. Yields of each of the mutants were similar suggesting the mutations were not detrimental to protein folding. These results further support *B. subtilis* as an excellent expression host for the production of proteases and other proteins that are difficult to produce in *E. coli*.

Figure 4.3 Purification of a typical His-tagged mutant of SGT from *B. subtilis* culture using Talon metal affinity resin. Lane A, 1 mL supernatant; Lane B, 1 mL column flow through; Lane C, 1 mL wash buffer. Lane D: concentrated recombinant protein (2 μg). Samples were concentrated using trichloroacetic acid in lanes A-C.

4.3.2 Techniques for Characterization of Substrate Specificity of Serine Proteases

Characterization of the substrate specificity of wild-type FXa using comprehensive libraries of small peptide substrates revealed the selectivity of the protease is not strict [41,42]. Interestingly, the P2 preference for large, planar hydrophobic substrates over Gly side chains has been noted in these studies. Moreover, P3 and P4 side chains are weakly selected for. On the basis of this specificity, FXa is thought to resemble a low efficiency trypsin rather than a highly selective thrombin [41]. One drawback of the libraries used to analyze the substrate specificity is the potential error caused by differences in conformation of each of the peptides in which each peptide may adopt differing conformations and lead to bias.
In the present study, we have used commercial substrates that differ from the natural polypeptide substrates or short synthetic peptides composed of L-amino acids. These substrates show enhanced specificity for members of the coagulation cascade and have been widely used in clinical applications. Non-standard amino acids are present including optimal blocking groups at their N-termini to generate highly specific substrates (Figure 4.4). Comparison of the rates of hydrolysis of these substrates shows that the mutations created in SGT have altered the active site geometry and substrate specificity of the enzyme.

4.3.3 Extended Substrate Specificity of SGT

With the T190P mutant of SGT as the starting point, extended substrate specificity at the S2 to S4 positions was introduced by substitution of residues found in FXa. The T190P mutant was chosen over the T190A mutation for the higher $K_m$ values. As the $K_m$ value for most peptides bearing the T190P mutant was approximately 50 to 100 µM, chromogenic substrates would be useful in the kinetic analysis of mutants. In contrast, the T190A mutant of SGT displayed $K_m$ values below 10 µM for each substrate. Kinetic dissection of subsequent mutants based on the T190A mutation would require the use of fluorescent substrates that are less commonly used for coagulation proteases. Moreover, the T190P mutant permitted estimation of the specificity constant ($k_{cat}/K_m$) using a direct interpretation of the rate of hydrolysis of peptide substrate. In this method, a single substrate concentration that is well below the $K_m$ value is hydrolyzed and the rate measured by spectrophotometry. The natural logarithm of substrate remaining as function of time is plotted as a straight line whose slope approximates $k_{cat}/K_m$. A number of mutants of SGT were characterized by this method and it was invaluable for determining successful alterations of substrate specificity.
Figure 4.4 Substrates used to characterize mutants of SGT with altered substrate specificity (Diapharma, West Chester, Ohio). Non-standard amino acids are incorporated into the substrate for improved discrimination among proteases.
4.3.4 Effect of the 99-loop on the Substrate Specificity of SGT

Substrate specificity of the S2 binding pocket in coagulation proteases is affected by the presence of a two or three amino acid insertion termed the 99-loop. In FXa, the insertion of two residues facilitates positioning of Tyr99 and restricts access to the S2 pocket. Characterization of substrate selectivity of SGT revealed preferences inherent within the protease (Figure 4.5). Spectrozyme PCa and S-2366 were preferred two-fold over all other substrates utilized. Both substrates have Pro at P2 and this likely leads to presentation of the Arg side chain in a conformation more favorable for hydrolysis. Importantly, the FXa preferred substrates were not preferred by the T190P mutant of SGT. Two mutants were designed to show the importance of this loop in generating selectivity in the S2 pocket (Figure 4.5). First, two residues, Lys96 and Glu97, were introduced to lengthen the 99-loop in SGT (denoted “LP bSGT”). All substrates examined were hydrolyzed with highly similar reaction rates to the T190P mutant using LP bSGT. As this mutant presented the smaller, hydrophilic side chain Thr at position 99, no selectivity in the S2 binding pocket was anticipated. Both of the inserted residues should orient their carbonyl oxygens into the S3/S4 pocket if the loop has the same conformation observed in other proteases. Introduction of T99Y led to a non-specific protease having similar preference for all substrates. These results suggested that the 99-loop was in a similar conformation to that observed in FXa and that additional determinants of substrate specificity were needed to reconstitute the desired specificity. Notably, substrates bearing P2 Gly were more effectively hydrolyzed suggesting a similar conformation of the loop to FXa.

Structural similarity of the FXa 99-loop is observed in aPC. This likely explains the maintained preference for aPC preferred substrates (Spectrozyme PCa and S-2366) in each of
the mutants characterized, including the mutant LP of SGT. In aPC, residue 99 is also Thr. Mutation of this position in aPC with substitutions found in FXa led to a protease with a similar a substrate specificity and inhibitory profile to FXa. In the same study, the opposite mutations in FXa (Y99T) led to a similar switching of specificity [140]. Importantly, as determined through peptide substrates, substrate specificity did not correlate with the hydrolysis of macromolecular substrates and demonstrates the crucial role for additional protein-protein interactions in the coagulation proteases.

![Graph showing normalized $k_{cat}/K_m$ values for the T190P, LP and YP mutants of SGT. Values derived from independent measurements done in triplicate (± 10% S.D.). Weak preference for aPC substrates is exhibited by the enzyme initially. Introduction of the two residue loop and T99Y generates a broadly specific protease.](image)

**Figure 4.5** Normalized $k_{cat}/K_m$ values for the T190P, LP and YP mutants of SGT. Values derived from independent measurements done in triplicate (± 10% S.D.). Weak preference for aPC substrates is exhibited by the enzyme initially. Introduction of the two residue loop and T99Y generates a broadly specific protease.

A number of other serine proteases have insertions at position 99. Large insertions (greater than 5 residues) are found in several kallikreins, the C1s protease, as well as complement factor B [144,145]. Compared to SGT, other proteases have a two or three residue insertion at this region including thrombin and pancreatic elastase II. In each of these enzymes, the loop plays a role in determining the substrate specificity of the enzyme of both the S2 and S3 substrate binding pockets. Increasing the length of the insertion correlates with
decreasing catalytic efficiency [145]. Futures studies for engineering substrate specificity should apply variable lengths of amino acid insertions and compositions of the 99-loop.

4.3.5 Mechanisms of P3 Selectivity in S1 Family Peptidases

Selectivity for P3 residues is poor in nearly all S1 peptidases. The enzyme-substrate interaction is limited due to solvent exposure of the P3 side chain. In FXa, P3 binding is generated by the side chains of residues 192 and 215. Throughout the entire family of S1 peptidases, residue 215 is highly conserved as a large planar side chain (W, F, Y). All crystal structures determined to date have shown the side chain in a conformation that borders the S4 specificity pocket (Figure 4.6). However, the backbone carbonyl group of residue 215 is involved in a hydrogen bond with the P3 residue of the substrate. In the crystal structures of wild-type recombinant bSGT and T190P mutant, residue 192 displayed a high degree of flexibility and was modeled as two conformations. Flexibility of this residue has been demonstrated in several coagulation proteases and the alternate conformations facilitate interactions with the P3 and P2' specificity pockets (Figure 4.7). As both SGT and FXa both possess Gln residues at position 192, mutation was not required. However, future work should involve mutagenesis of this residue due to its involvement in substrate specificity and inhibition by protease inhibitors.

Residue 192 has evolved to not disrupt substrate binding and plays a role in enzyme inhibition. Studies involving protein C, FXa, and thrombin have shown that residue 192 is important in protease-inhibitor contacts and is a key basis for differential inhibition [146-148]. Met/Glu/Gln substitutions at position 192 in coagulation factor IXa did not significantly alter substrate specificity using peptide substrates similar to those used in this study [143].
Presumably, mutation of the residue to a Lys side chain in SGT would generate an increased preference for acidic side chains at P3. Unfortunately, Lys192 has not been introduced into any S1 peptidases by site-directed mutagenesis yet the residue exists naturally in some members of the family.

Figure 4.6 S3 & S4 binding pockets of FXa. Gln192 plays a limited role in determining the selectivity of the S3 pocket stemming from inherent flexibility of the side chain. In the FXa structure without a substrate, the side chain points away from S3 pocket. The model depicted is the superimposition of FXa (PDB ID 1HCG) and FVIIa in complex with 1,5-dansyl-Glu-Gly-Arg-chloromethyl ketone (PDB ID 1CVW), with the FVIIa structure hidden.
Figure 4.7 S3 and S4 binding pockets of FVIIa. In the crystal structure of human coagulation factor VIIa in complex with 1,5-dansyl-Glu-Gly-Arg-chloromethyl ketone (PDB ID 1CVW), Lys192 interacts with the negatively charged side chain in P3 of the inhibitor.

Sequence analysis of the S1 peptidases reveals a subset of proteases that contain positively charged side chains at position 192. These proteases show S3 selectivity for acidic P3 residues. Notably, venombin A and bilineobin from the moccasin snake (*Agkistrodon bilineatus*) have Lys192 [149,150]. These proteases are found in the venom of the snake and
mimic FXa by activating thrombin at the same position in the polypeptide. Although crystal structures of these proteases have not been reported, molecular modeling of bilineobin suggests that residue 192 occupies an identical position to that observed in FXa, thrombin and SGT [151]. Creation of an electrostatic interaction between the substrate and enzyme may decrease the rate of deacylation during catalysis. Rates of substrate hydrolysis by Lys192 bearing proteases are 10-fold lower than observed for the coagulation proteases. Thus, the Q192K mutant of SGT should be characterized and may yield a stronger preference for acidic side chains at P3 with a concomitant reduction in catalytic efficiency. However, it is likely that the high flexibility of residue 192 will limit the maximal stringency for P3 side chains.

Two mechanisms have evolved to increase the selectivity of the S3 pocket in the S1 family of peptidases. In rat mast cell protease II, the absence of a disulphide bond between residues 191 and 220 generates additional enzyme-substrate contacts and enlarges the S3 pocket [152]. P3 selectivity can also be generated by a secondary protein as evident in the structure of staphylokinase in complex with the protease domain of plasmin. Staphylokinase acts as a co-factor and inserts several side chains into the S3/S4 pocket. As a result the substrate specificity and inhibitory profile of plasmin are drastically altered [153]. Thus, the serine protease scaffold is highly amenable for further engineering of selectivity in the S3 binding pocket.

4.3.6 Role of the 172-loop and Residue 217 in the S1 Peptidases

In order to form the S4 pocket in FXa, a stretch of residues from 172 to 174 adopt a conformation distinct from that observed in broadly specific trypsin like enzymes (Figure 4.8). In SGT and other non-specific serine proteases, a large hydrophobic side chain at residue
172 buries itself into core of the enzyme. In turn, the backbone of residues 173 and 174 adopts a conformation that exposes the side chains of these residues away from the S4 pocket. In FXa, FIXa, thrombin and aPC residue 172 is Thr, Met, or Ser. In these proteases the 172-loop adopts an “up” conformation and bounds the S4 pocket. In order to generate the proper conformation of the 172-loop, two mutations were made in SGT: Y172S and N174F.

**Figure 4.8** Conformation of the 172-loop in S1 peptidases: SGT (A), FXa (B), aPC (C) and FVIIa (D). In FXa and aPC, the 172-loop adopts a conformation such that residue 172 is not buried in the core of the enzyme which is evident in all structures of broad specificity trypsin-like enzymes. FVIIa has a large insertion at residue 170, and generates substrate specificity in a different manner than other coagulation proteases.
Introduction of N174F alone into SGT yielded a protease with selectivity for S-2366, a substrate designed for quantification of FXIa and aPC. Based on the crystal structure of T190P SGT, N174F should not affect the specificity of the S4 pocket unless the conformation of the 172-loop is distorted. The alteration of substrate specificity may be due to the Phe side chain at position 174 replacing residue Tyr172 its buried conformation. Subsequent mutagenesis of Y172S suggested that Phe174 was in a buried conformation as it resulted in a very little change in substrate specificity compared to the YFP mutant of SGT (Figure 4.9). A recent study reported the crystal structure of rat anionic trypsin with similar mutations in the 172-loop, and suggests the basis for aPC and FXIa-like specificity [142].

In the crystal structure of a mutant of rat anionic trypsin with the Ser-Ser-Phe sequence substituted at the 172 to 174 positions, the 172 loop adopts a novel conformation in which Phe174 buries inward in the structure of the enzyme similar to that anticipated in the mutants of SGT [142]. As a result, the conformation of the side chains at positions 172 and 173 do not mimic those observed in FXa even though the polypeptide sequence is the same. The S4 pocket is enlarged as the 172-loop extends farther away from the structure. If a similar conformation occurred in the YFP and YSFP mutants of SGT then the preference for substrates with larger side chains at P4 might be more favored. Both S-2366 and S-2266 have large hydrophobic groups at P4 (pyroglutamic acid and D-valine, respectively).
Figure 4.9 Normalized $k_{cat}/K_m$ values for the YFP and YSFP mutants of SGT.
Introduction of Y172S yielded no substantial change in substrate specificity. Values derived from independent measurements done in triplicate (± 10% S.D.).

Addition of the E180M mutation to the YSFP SGT mutant (YSFMP) led to significant improvement in FXa-like specificity (Table 4.3). In any known crystal structure of an S1 family peptidase, the Met side chain at residue 180 makes no direct contacts with the substrate or any component of the protein. A limited diversity of amino acid variation is observed at position 180 in the S1 peptidases with a Met residue present in ~60% of the all proteins in the family. Only a few of S1 peptidases bear a positively charged side chain, such as Lys or Arg, at this position. Positioning of residue 180 is achieved through a β-hairpin formed by residues 177 to 180. The NH group of residue 180 forms a hydrogen bond with the C=O backbone of residue 177. In SGT, the carboxylate moiety of Glu180 also forms a hydrogen bond with the NH group of the amide backbone of Val177. Mutation of residue 180 should alter the electrostatic environment of the S3 pocket as well as permit an alternate conformation of the 172-loop.

Dissection of the kinetic constants of the YSFMP mutant bearing the E180M mutation indicates moderate reconstitution of FXa-like properties. Boc-Leu-Gly-Arg was the most
specific substrate for the mutant protease indicating the S3 pocket is hydrophobic in the mutant protease. Substrates designed for the quantification of FXa had the lowest \( K_m \) values of all substrates (S-2222 and Spectrozyme FXa). However, substrates with the lowest \( K_m \) values were not associated with the highest turnover numbers. Efficient turnover of non-FXa preferred substrates still occurred suggesting the extended binding pocket of FXa was not present. Based on the observed properties of the YFP, YSFP and YSFMP mutants of SGT, stabilization of the 172-loop was implicated as being necessary for generating FXa-like extended substrate specificity. Further, crystallographic analysis is required to observe the conformation of the 172-loop caused by these mutations.

<table>
<thead>
<tr>
<th>Substrate</th>
<th>( K_m ) (( \mu \text{M} ))</th>
<th>( k_{\text{cat}} ) (s(^{-1} ))</th>
<th>( k_{\text{cat}}/K_m ) (( \mu \text{M}^{-1} \text{s} ))</th>
<th>FXa ( k_{\text{cat}}/K_m ) (( \mu \text{M}^{-1} \text{s} ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>LGR</td>
<td>460 ± 12</td>
<td>4664 ± 811</td>
<td>10.1</td>
<td>-</td>
</tr>
<tr>
<td>Boc-Leu-Gly-Arg-pNA</td>
<td>460 ± 12</td>
<td>4664 ± 811</td>
<td>10.1</td>
<td>-</td>
</tr>
<tr>
<td>S-2222</td>
<td>241 ± 36</td>
<td>1765 ± 959</td>
<td>7.3</td>
<td>4.9</td>
</tr>
<tr>
<td>Bz-Ile-Glu(( \gamma )-OR)-Gly-Arg-pNA (R = H or CH(_3) at 50%)</td>
<td>241 ± 36</td>
<td>1765 ± 959</td>
<td>7.3</td>
<td>4.9</td>
</tr>
<tr>
<td>S-2366</td>
<td>511 ± 67</td>
<td>3683 ± 1209</td>
<td>7.2</td>
<td>-</td>
</tr>
<tr>
<td>pyroGlu-Pro-Arg-pNA-HCl</td>
<td>511 ± 67</td>
<td>3683 ± 1209</td>
<td>7.2</td>
<td>-</td>
</tr>
<tr>
<td>Spectrozyme PCa</td>
<td>623 ± 139</td>
<td>3788 ± 620</td>
<td>6.1</td>
<td>4.1</td>
</tr>
<tr>
<td>H-D-Lys(g-Cbo)-Pro-Arg-pNA</td>
<td>623 ± 139</td>
<td>3788 ± 620</td>
<td>6.1</td>
<td>4.1</td>
</tr>
<tr>
<td>Spectrozyme FXa</td>
<td>356 ± 20</td>
<td>1800 ± 240</td>
<td>5.1</td>
<td>18.2</td>
</tr>
<tr>
<td>M-D-CHG-Gly-Arg-pNA</td>
<td>356 ± 20</td>
<td>1800 ± 240</td>
<td>5.1</td>
<td>18.2</td>
</tr>
<tr>
<td>S-2266</td>
<td>2271 ± 468</td>
<td>4017 ± 1687</td>
<td>1.8</td>
<td>-</td>
</tr>
<tr>
<td>H-D-Val-Leu-Arg-pNA</td>
<td>2271 ± 468</td>
<td>4017 ± 1687</td>
<td>1.8</td>
<td>-</td>
</tr>
<tr>
<td>S-2238</td>
<td>1396 ± 80</td>
<td>1056 ± 398</td>
<td>0.8</td>
<td>1.2</td>
</tr>
<tr>
<td>H-D-Phe-Pip-Arg-pNA</td>
<td>1396 ± 80</td>
<td>1056 ± 398</td>
<td>0.8</td>
<td>1.2</td>
</tr>
<tr>
<td>S-2302</td>
<td>1467 ± 80</td>
<td>512 ± 220</td>
<td>0.3</td>
<td>-</td>
</tr>
<tr>
<td>H-D-Pro-Phe-Arg-pNA</td>
<td>1467 ± 80</td>
<td>512 ± 220</td>
<td>0.3</td>
<td>-</td>
</tr>
<tr>
<td>S-2251</td>
<td>1857 ± 103</td>
<td>527 ± 260</td>
<td>0.3</td>
<td>-</td>
</tr>
<tr>
<td>H-D-Val-Leu-Lys-pNA</td>
<td>1857 ± 103</td>
<td>527 ± 260</td>
<td>0.3</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 4.3 Steady-state kinetic parameters for the hydrolysis of a series of p-nitroanilide chromogenic substrates by the YSFMP mutant of SGT compared to that observed with FXa under similar reaction conditions. (FXa data from ref. [140]). Values obtained in triplicate (± S.D.).
Within the coagulation proteases, residue 217 plays a role in substrate selectivity. Mutations at this position have been characterized in thrombin, coagulation factor VIIa (FVIIa), and coagulation factor IXa (FIXa) [143, 154, 155]. In these studies, residue 217 was described as a determinant of P2/P3 selectivity via formation of direct contact with the substrate. Conclusions drawn were based on analysis of substrate specificity using only a few peptide substrates and primarily through enzyme-inhibitor interactions evident in crystal structures. The latter can lead to misrepresentation of the true function of peripheral residues of the active site, such as residue 217, as the interaction between enzyme and inhibitor is typically far tighter than for a natural substrate. Moreover, later studies established the significance of a sodium binding site in several of the coagulation factor proteases. Binding of a sodium ion improves the catalytic efficiency of the enzyme by structural changes in the protease domain. Residue 217 has been implicated in stabilization of the ion binding site [156]. Inspection of crystal structures of FXa, VIIa, and thrombin in complex with peptide based inhibitors supports neither of these theories completely [141, 157, 158]. Residue 217 is typically a negatively charged side-chain throughout the S1 family of peptidases, yet this residue is a Tyr in SGT. In several crystal structures of coagulation proteases, the carboxyl group of the Glu side chain is involved in the formation of two charge assisted hydrogen bonds with residue 173 in the 172-loop. Hence, it was postulated that introduction of this side chain would lead to stabilization of the loop in an upwards conformation and enhance FXa-like extended substrate specificity.

Kinetic analysis of the YSFMPE mutant of SGT bearing the Glu217 yielded a protease more similar to FXIa and not FXa (Table 4.4). Little is known about the substrate specificity of FXIa in vitro and no crystal structure has been reported for the protein. Inspection of the
polypeptide sequence of human FXIa shows the presence of Met180 and Glu217, and a three residue insertion loop at position 99 that would present either Ser or Gly into the S2 pocket [159]. Crystallographic analysis is required to determine whether the introduced Glu217 side chain adopts a conformation that restricts the P4 pocket or whether it forms an electrostatic interaction with Arg222. As in the YSFP and YSFMP mutations, the Tyr172 in the polypeptide sequence of FXIa may possess a 172-loop in the down conformation. Hence, all of the determinants thought to generate FXa-like specificity are present in FXIa with the exception of Tyr99 and may explain substrate specificity of the YSFMPE mutant of SGT.

<table>
<thead>
<tr>
<th>Substrate</th>
<th>$K_m$ (µM)</th>
<th>$k_{cat}$ (min$^{-1}$)</th>
<th>$k_{cat}/K_m$ (µM$^{-1}$ min$^{-1}$)</th>
<th>FXa $k_{cat}/K_m$ (µM$^{-1}$ min$^{-1}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>S-2366 pyroGlu-Pro-Arg-pNA-HCl</td>
<td>526 ± 23</td>
<td>6691 ± 1409</td>
<td>12.7</td>
<td>-</td>
</tr>
<tr>
<td>S-2302 H-D-Pro-Phe-Arg-pNA</td>
<td>568 ± 39</td>
<td>5487 ± 1569</td>
<td>9.7</td>
<td>-</td>
</tr>
<tr>
<td>Spectrozyme FXa M-D-CHG-Gly-Arg-pNA</td>
<td>167 ± 37</td>
<td>819 ± 180</td>
<td>4.9</td>
<td>18.2</td>
</tr>
<tr>
<td>S-2266 H-D-Val-Leu-Arg-pNA</td>
<td>488 ± 33</td>
<td>1624 ± 343</td>
<td>3.3</td>
<td>-</td>
</tr>
<tr>
<td>Spectrozyme PCa H-D-Lys(g-Cbo)-Pro-Arg-pNA</td>
<td>1952 ± 34</td>
<td>5870 ± 1553</td>
<td>3.0</td>
<td>4.1</td>
</tr>
<tr>
<td>LGR Boc-Leu-Gly-Arg-pNA</td>
<td>177 ± 22</td>
<td>196 ± 86</td>
<td>1.1</td>
<td>-</td>
</tr>
<tr>
<td>S-2238 H-D-Phe-Pip-Arg-pNA</td>
<td>316 ± 39</td>
<td>198 ± 86</td>
<td>0.6</td>
<td>1.2</td>
</tr>
<tr>
<td>S-2222 Bz-Ile-Glu(y-OR)-Gly-Arg-pNA (R = H or CH$_3$ at 50%)</td>
<td>2341 ± 126</td>
<td>1066 ± 701</td>
<td>0.5</td>
<td>4.9</td>
</tr>
</tbody>
</table>

Table 4.4 Steady-state kinetic parameters for the hydrolysis of a series of p-nitroanilide chromogenic substrates by the YSFMPE mutant of SGT compared to that observed with FXa under similar reaction conditions. (FXa data from ref. [140]).

Values obtained in triplicate ± S.D.
4.3.7 Additional Elements Needed for Reconstructing FXa-like Specificity

Additional mutations must be made to improve further FXa-like selectivity of the SGT mutants constructed in the present study. Non-additive effects of mutations are common for mutations that are residues in close proximity [153,160,161]. Hence, the absence of one specificity determinant may hinder the ability of other residues to function properly. Studies attempting reconstitution of FXa specificity in other proteases may yield clues as to what element is absent from the YSFMPE mutant of SGT. Unfortunately, these studies for the most part have focused exclusively on the 99-loop [70].

Rat anionic trypsin (RAT) was mutated in several of the regions characterized in this study but did not lead to an enzyme with FXa-like properties [142]. Mutations of residue 190, the 99-loop, and the 172-loop were combined in RAT but were characterized on the basis of enzyme inhibition rather than substrate specificity. Inhibition constants (K_i) of inhibitors of FXa were typically 10-fold higher with the combined mutant of RAT. On the basis of structure data, certain inhibitors could assist the stabilization the 172-loop in the proper upwards conformation. Based on the data from SGT mutants in the present study, it appears that the conformation of the 172-loop is the key ingredient missing for reconstitution of the extended substrate specificity of FXa.

Several possibilities exist for further stabilization of the 172-loop. Introduction of Y217E was unsuccessful at improving the selectivity of the YSFMP mutant of SGT and may be due to the requirement of a hydroxyl group at the side chain of residue 173, which is a Gly in SGT. Residues within a 6 Å radius of residues 172 to 174 in FXa include residues 167 to 176, 182, 215 to 217, 224 and 227. Comparison of the crystal structures of SGT and FXa show that most of these residues are identical and exist in highly similar conformations in
both proteases. Substitutions of the residues adjacent to the 172-loop in the polypeptide sequence may further stabilize the loop. However, the largest differences in structure are evident at residues 224 and 227. Importantly, these residues are involved in the formation of the Na$^+$ binding site in FXa. It is possible that introduction of a Na$^+$ binding site into YSF MPE SGT will reconstruct the enzymatic properties of FXa. If the site were present, burial of residue 172 or 174 into the core of the enzyme could be disfavored. Importantly, the YSF MPE mutant has kinetic properties similar to FXIa which also does not possess a Na$^+$ binding site. Sequence analysis of the coagulation proteases indicates that in their evolutionary past, all other coagulation proteases had a Na$^+$ binding site.

Site specific sodium binding has been demonstrated to play a key role in the catalytic efficiency of the coagulation factor proteases including FXa, aPC, and thrombin. Di Cera demonstrated the crucial role of residue 225 in the serine proteases for binding a single sodium ion near the active site of the enzyme [162]. Binding of sodium and no other alkali metal to these enzymes generates a 3- to 5-fold increase in catalytic efficiency. In nearly all serine proteases, residue 225 is occupied by a Pro and these proteases do not demonstrate increased catalytic activity in the presence of sodium ions. However, in the coagulation factor proteases, this residue is Tyr. Absence of Pro at position 225 allows proper positioning of the carbonyl group of the preceding residue in the polypeptide sequence to bind the Na$^+$ ion [77]. In addition to residue 225, an extensive network of hydrogen bonds and water molecules facilitates stabilization of a sodium ion in a position that is very near the S1 binding pocket (Figure 4.10). Much research has been devoted to the characterization of the ion binding site in thrombin, activated protein C and coagulation factor Xa [156,163,164]. Disruption of any component of the ion binding site readily abolishes binding and catalytic activity. The
complexity of these enzymes has limited our ability to understand the structural and thermodynamic effects of sodium binding. However, the crystal structure of thrombin in the absence of Na$^+$ suggests that removal of the ion destabilizes the entire S1 pocket and changes the conformation of the Cys168-Cys182 disulfide bond [165]. The S1 pocket lies adjacent to residue 172 in SGT or residue 174 in Y172S N174F mutants of SGT in the buried conformation. Therefore, the conformation of the loop could be altered by the presence of an occupied Na$^+$-binding site.

**Figure 4.10** Na$^+$-binding site in thrombin. Octahedral co-ordination stabilizes the ion. Backbone carbonyl groups and water molecules are involved in the interaction. Three ion-pairs surround the site and provide stability.

Introduction of a sodium binding site into a mutant SGT to improve substrate specificity would involve significant mutagenesis of the gene as nearly all of the residues that
comprise the site are absent in SGT. In thrombin, the Na$^+$ binding site is located between two surface loops beginning at residues 180 and 220 [77]. The ion binding site is 15-20 Å distal from the catalytic triad, yet lies within 5 Å from D189. A cylindrical cavity occupied by up to sixteen water molecules helps to stabilize the site. Bound Na$^+$ is coordinated octahedrally by two carbonyl oxygen atoms provided by R221a and K224, and four buried water molecules. One of these water molecules hydrogen bonds to the side chain of D189 establishing a direct link between the sodium ion and the S1 site [52]. Overall stability of the Na$^+$ site is provided by three ion pairs, R221a-E146, K224-E217, and D222-R187. Sodium binding involves residues adjacent to the substrate binding pocket and may play a role in generating stringent substrate specificity. In particular, residues 192 and 217 were both targets for mutagenesis in the present study and have been implicated in sodium binding.

### 4.3.8 Utility of a FXa-like Protease

Factor Xa is routinely used for the cleavage of recombinant fusion proteins, yet the preference for the IEGR cleavage sequence is not strict. For example, in honey bee prepromelittin, a related sequence, VLGR, was readily cleaved by FXa [166]. Although this situation is not common and can be prevented by inspection of the polypeptide sequence of the recombinant protein, a more specific protease is desirable. A number of vendors supply alternative proteases with differing recognition sequences, including TEV protease [167], thrombin [127], enterokinase [127], and PreScission$^\text{TM}$ protease [127] (Table 4.5). However, many plasmids currently employed contain the FXa cleavage motif in-frame with common restriction sites for cloning and recombinant expression of proteins. To assess the ability of mutant SGT proteases to hydrolyze the bond after the Ile-Glu-Gly-Arg sequence in
macromolecular substrates, prothrombin was digested with limited amounts of each mutant constructed in SGT (Figure 4.11). Prothrombin was chosen as it is the substrate of FXa in vivo and has a number of potential cleavage sites.

<table>
<thead>
<tr>
<th>Protease</th>
<th>Cleavage Sequence</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enterokinase</td>
<td>Asp-Asp-Asp-Asp-Lys</td>
<td>P1' can not be a Pro.</td>
</tr>
<tr>
<td>Factor Xa</td>
<td>Ile-Glu/Asp-Gly-Arg</td>
<td>Secondary cleavage sites (due to low P3 and P4 preference)</td>
</tr>
<tr>
<td>Thrombin</td>
<td>Leu-Val-Pro-Arg(^{+})Gly-Ser</td>
<td>Secondary cleavage sites. Considerably more expensive than FXa</td>
</tr>
<tr>
<td>TEV protease</td>
<td>Glu-Asn-Leu-Tyr-Phe-Gln(^{+})Gly</td>
<td>Less useful for removal of C-terminal tags</td>
</tr>
<tr>
<td>PreScission™protease</td>
<td>Leu-Glu-Val-Leu-Phe-Gln(^{+})Gly-Pro</td>
<td>Remaining sequence after hydrolysis is problematic</td>
</tr>
</tbody>
</table>

Table 4.5 Cleavage sites of proteases used in processing recombinant proteins.

Accurate processing of prothrombin was not achieved by any of the mutants constructed in this study. Overnight digestions are typically used for site specific proteolysis of recombinant fusion proteins and hence a similar strategy was employed. Close inspection of the digestion pattern shown in Figure 4.11 shows that only YSFMP SGT yields a proteolytic fragment of similar size to the B-chain of thrombin. However, additional cleavages occur and leave only trace amounts of the desired fragment. The YSFMPE mutant of SGT did not process prothrombin to yield any fragment suggestive of FXa-like specificity and confirms the data provided through hydrolysis of small peptide substrates. Prothrombin activation is a poor representative of what would be anticipated for the cleavage of typical recombinant proteins. In particular, several regions of the protein are optimized for hydrolysis.
Therefore, it is likely that the YSFMP mutant of SGT could be used as an alternative for the site specific proteolysis of recombinant proteins as an alternative to FXa.

**Figure 4.11** Prothrombin processing by mutants of SGT: T190P (B), YP (C), YSFP (D), YSFMP (E) and YSFMPE (F) mutants of SGT compared to FXa (G) and undigested prothrombin (A). Only the YSFMP mutant of SGT processed trace amounts of a product the same size as thrombin. The YSFMPE mutant of SGT contains all known specificity determinants of FXa, but did not generate any of the fragments associated with the activation of prothrombin.
4.4 Conclusions & Future Directions

Extended substrate specificity in FXa is the result of four amino acids at positions 99, 174, 180, and 192. These residues are positioned accurately by residues 172 and 217, and a two amino acid insertion at position 99. Other residues that surround these positions are likely involved in minor optimization of the electrostatic environment and stability of the region. Substitution of the key residues of FXa into SGT created a protease with similar substrate specificity that was more similar to FXIa rather than FXa. Based on these findings, the development of a protease with more stringent specificity for the preferred FXa cleavage sequence may be possible through additional mutation of SGT particularly by addition of the $\text{Na}^+$ binding site as found in FXa or thrombin. Central to the success of this research was the use of *B. subtilis* for production of the recombinant protein and mutants thereof. The ease of protein purification and low cost of production is a significant advantage over previously reported systems. Structural and sequence similarity of SGT to FXa suggests that substrate specificity of any coagulation protease could be re-created if not bettered.

Increased stringency for the Ile-Glu-Gly-Arg FXa could be generated by mutagenesis of the 99-loop and position 192 in the polypeptide sequence of the YSFMPPE mutant of bSGT. Characterization of the specificity of FXa shows a weak preference at the P2, P3 and P4 positions of the substrate. Optimization of the 99-loop created in SGT will be needed to decrease flexibility and restrict access to the S2 pocket. However, it is unknown what mutations will create rigidity in the 99-loop. Mutagenesis of residue 192 to a Lys amino acid should increase the specificity of the S3 pocket for negatively charged side chains. Increased stringency in the S4 pocket may be created by strengthening the interaction of residues 173 and 217. In particular, mutants bearing Ser, Thr, or Lys at position 192 may stabilize the loop
further. Each of these mutations will likely cause a decrease in the catalytic efficiency of the enzyme. The suggested amino acids are likely not observed in the wild-type FXa protein due to the physiological requirement of efficiency. Evolution has decreased the potential harmful effect of less than perfect substrate specificity through linkage of additional protein domains that facilitate protein-protein and protein-lipid interactions.
Chapter 5. General Discussion and Outlook

5.1 Substrate Specificity Determinants of the S1 family of Serine Proteases

Increasing the substrate specificity through mutagenesis of the S1 to S4 pockets in SGT was successful. Six mutations were combined, a two residue insertion and five point mutations, to mimic the active site architecture of coagulation factor Xa. Only the combined mutant demonstrated a strong preference for the desired Ile-Glu-Gly-Arg recognition sequence. Introduction of a seventh mutation in SGT, Y217E, did not further improve the specificity towards FXa preferred substrates. As anticipated, specificity results largely from the amino acid residues which constitute the enzyme-substrate interface but additional regions of the protease are important. A further increase in specificity will require stabilization of the active site architecture and optimization of the electrostatic environment of the entire active site. These mutations will involve second shell or more distal residues throughout the protein. Molecular evolution over millions of years has accomplished these tasks and has generated diverse proteases and substrate specificities.

A number of proteases share similar architecture and substrate specificity determinants as SGT. In the MEROPS database, SGT is classified as a member of Clan SA, Family S1, Subfamily A peptidases [168]. Nearly 1000 proteases in this family have been identified, including the vertebrate and invertebrate coagulation factor proteases, kallikreins, granzymes, and, complement proteases (Figure 5.1). Notably, these proteases demonstrate trypsin-like, chymotrypsin-like, and elastase-like primary substrate specificities (Table 5.1). Although the overall architecture of these proteases is similar, variations in the active site of the enzyme facilitate differing specificities.
Figure 5.1 Simplified phylogenetic tree of the S1 family of peptidases. In each of the S1 sub-families of peptidases, a diversity of substrate specificities are found. Subfamily A is considerably larger than the other sub-families which are limited in distribution to gram positive/negative bacteria and viruses.
Table 5.1 Substrate specificities of S1 family peptidases.

Chymotrypsin-like proteases require proper positioning of large hydrophobic side chains in the primary binding pocket for efficient catalysis. Hedstrom demonstrated that conversion of a trypsin-like enzyme to a chymotrypsin-like enzyme required extensive mutagenesis of the S1 pocket [56,57,59]. In addition to point mutations at positions 189, 216, and 226 several loops adjacent to the pocket were required to generate the change in the primary substrate specificity but resulted in a poorly active enzyme. Importantly, the altered loops do not contact the substrate directly. Improvement of the catalytic efficiency against amide substrates was achieved by mutagenesis of Y172W [57]. In the coagulation proteases, the side chain of residue 172 is exposed to the solvent and involved in the S4 substrate binding pocket [178]. In the present study, mutation of residue 172 in SGT was required to mimic FXa-like specificity of the S4 pocket. However, burial of the aromatic side chain at position led to an alternate conformation in the 172-loop. These studies demonstrate that multiple amino acids act in concert and that similar positions in the polypeptide sequence can affect differing specificity pockets in the folded protein.
Amino acid insertions in the protein sequence are a major component in the substrate specificity of serine proteases. Hydrogen bonds, dipole moments, steric constraints, and altered electrostatic environments can be generated by the introduction of one or more residues into the protein. In the present work, addition of a two residue loop at position 99 was important in altering the properties of the S2 pocket. Loop insertions are found in many members of the S1 family of proteases.

In addition to position 99, loop insertions are found at other positions in the protease sequence in the S1 family peptidases. Thrombin possesses a 10 amino acid insertion at position 60 as well as a shorter loop at position 148 that are involved in substrate recognition [179,180]. These loops have limited flexibility that has only been demonstrated by mutagenesis of the protease [181]. Other positions amenable for loop insertions include residue 170 found in several coagulation proteases [182], the “kallikrein loop” at position 90 [183], and also at residue 70 in the neuropsins [184]. The loops can also be involved in biochemical properties other than substrate specificity such as enzyme regulation, protein-protein interactions and zymogen activation.

Extended substrate specificity within the S1 family of peptidases can involve residues on the N- and C-termini of the scissile bond of the substrate. Vertebrate coagulation proteases possess substrate specificity in the S1 to S4 pockets. Small side chains at P1’ are also favored by many proteases due to their decreased steric hindrance with the enzyme. Members of the kallikrein family display selectivity at P2’ for Arg side chains. Rat and human mast cell proteases have a marked preference for acidic residues at P2’ due to the presence of Lys40, Arg143, and Lys192 [185]. Hence, it seems possible that a similar strategy employed to
generate coagulation factor Xa specificity in SGT could be applied to generate proteases highly selective for amino acids on both sides of the scissile bond.

5.2 Molecular Evolution of the S1 family of Serine Proteases

Over the course of evolution, proteases have been incorporated into a wide variety of cellular processes. Although the protease domain provides the catalytic machinery, additional protein domains are linked to add functionality. Substrate recognition and cellular (or extracellular) localization are two common roles of these domains. For example, the CUB domains facilitate protein-protein interactions [186], Gla domains promote protein-lipid interactions [187] and fibronectin domains localize proteins to extracellular fibrin depositions [188,189]. Based on the diversity of associated domains, construction of an accurate phylogenetic tree is a difficult task.

A number of approaches have been taken to dissect the evolution of the serine proteases. These studies can be divided into those that examine the entire protease sequence including the associated protein domains, and those which examine only the protease domain or parts thereof [190-194]. All studies have supported the existence of a single ancestor for the entire S1 family of peptidases [120]. Approximately 40 amino acid positions are highly conserved throughout the family to produce a consistent three dimensional structure. Variations in these residues are limited to conservative mutations that preserve the polarity and size of the side chain [117]. The catalytic triad His-Asp-Ser is present in all members of the family [195]. A carboxyl-group containing amino acid at position 194 is also absolutely conserved amongst the family. Asp/Glu194 provides the electrostatic interaction with the N-terminus of the protease domain (residue 16) [196]. Formation of this interaction stabilizes the
oxyanion hole and the active site catalytic triad. Limited diversity at a particular residue and close proximity to the active site suggests involvement in enzyme-substrate interactions and specificity.

Analysis of the amino acid distribution at each residue in the substrate binding pockets shows a limited number of possibilities exist in nature (Table 5.2). Determinants characterized in the present work (residues 180, 190, and 217) are moderately conserved, and can be mutated further to yield novel proteases. Other active site residues with less sequence conservation (residues 215 and 228) have not been characterized with respect to substrate specificity but likely can be manipulated to influence specificity. Together with the variability in size and composition of the loops that surround the active site, the ability to design proteases with specificities that do not exist in nature seems possible. At present, it has not been established how highly specific proteases could be developed in the laboratory aside from structure based design. Elements of the present dissertation could be used in such a system.

5.3 Paper, Rock, Scissors Genetic Screening of Trypsin-like Proteases

Design of substrate specificity by structure based techniques is limited by our inability to understand the complex interactions that occur in a protein structure. Mutations thought to exhibit an effect often result in absent or opposing results to that expected [125,197]. In the present study, both T190P and Y217E mutations in SGT yielded significantly different kinetic properties from what were anticipated. Thus, alternate approaches are required. Randomized mutagenesis of the whole gene or parts thereof combined with genetic selection has been widely adopted for the addition or manipulation of enzymatic properties.
### Table 5.2 Substrate specificity determinants of S1 family sub-family A peptidases (not including loop regions).

Sequence alignment of 740 proteases in the S1 family peptidases shows a limited diversity of amino acids exists at the substrate binding region in all known members of the family (Sequence analysis described in Chapter 4.2.3). The final mutant construct (YSFMPE) did not yield the desired specificity, even though all known specificity determinants were incorporated.

Converting trypsin, or other primitive proteases, into highly specific enzymes suffers from the disadvantage that the wild-type enzyme will almost always be more active towards any substrate than the target enzyme. Typical directed evolution strategies employ creating novel activities or substrate specificities that are completely disparate from the wild-type enzyme [198,199]. For example, the wild-type enzyme does not catalyze a particular reaction.
or does so poorly and the screen selects for increased activity. Such a screen can not be
applied to increase proteolytic substrate specificity from a primitive enzyme. A genetic screen
for novel substrate specificity must also compensate for the large diversity of possible
substrates arising from 20 amino acids at each position. A three residue stretch of amino acids
can have 8000 different possible permutations. Moreover, the side chains have a degree of
similarity that can not easily be accounted for. Thus, identification of the desired mutant
protein must rely on strategies employed in nature. In particular, protease inhibitors may be
useful for influencing the molecular evolution of proteases.

Recent molecular evolution of HIV proteases serves as an excellent model for directed
evolution to escape inhibition. For the virus to reproduce, a number of proteins are produced
as precursors that must be cleaved for activation [200,201]. On the basis of this requirement, a
number of protease inhibitors have been designed and applied clinically in the treatment of
this disease [202]. Unfortunately, the virus is known to mutate rapidly and a number of
mutations have been described that directly lead to resistance against inhibition [203]. These
mutations are located at the active site cleft, as well as at distal regions of the protease.
Notably, resistance to inhibition produced altered substrate specificity of the enzyme [204].
Mutations in the protein substrates at the site of cleavage have been demonstrated to
accompany the mutations that generate inhibitor resistance [205]. These observations suggest
that evolution directed by inhibition is a valid concept for the design of highly specific
proteases.

Genetic screening for novel substrate specificity could be generated by combining
three components: a protease, an inhibitor that affects the wild-type but not the target
protease, and a means by which to visualize activity. Paper, Rock, Scissors genetic screening
is put forth to accomplish these tasks (Figure 5.2). Co-expression of an inhibitor with the protease will provide the direction for the evolution by distinguishing proteases that are similar to the wild-type in the active site. Variable sized loops at different positions in the polypeptide sequence of SGT could be added and the whole gene randomly mutated by various means [206,207]. Only active proteases with altered active site geometry or properties would overcome the inhibition and potentially yield mutants with improved substrate specificity. Detection of protease activity could apply a sensitive fluorescent substrate that is added directly to the solid growth medium. Formation of a halo would show proteolytic activity and signal a potentially important mutant that could be characterized further. Significant technical hurdles are evident for this system to function. In particular, the generation of large mutant libraries in \textit{B. subtilis} is significantly more difficult than \textit{E. coli} and no known protease substrate would be amenable for \textit{in vivo} selection. To overcome these limitations new techniques and novel proteins must be developed.

\textit{E. coli} trypsin inhibitor, ecotin, is a potent inhibitor of trypsin-like serine proteases with several properties amenable for use in directed evolution. The binding mode of ecotin is identical to that observed with a natural substrate [208]. Ecotin displays a number of useful properties, including thermostability and stability in extreme pH. The inhibitor exists as a dimer in solution and has two regions involved in inhibitor-protease interactions (Figure 5.3) [209]. Marked differences in inhibitory strengths are observed using the wild-type protein and high specificity enzymes (Table 5.3). Coagulation Factor Xa and thrombin have a 100- to 1000-fold lower inhibition constant compared to broadly specific proteases [210]. A more potent inhibitor against trypsin-like enzymes can be constructed through mutagenesis of the P1 Met to Arg [211]. In contrast, the inhibitor can be converted to a monomeric form with
1000-fold less inhibitory strength towards trypsin-like enzymes [132]. Thus, ecotin can be genetically manipulated to exhibit a broad range of inhibitory strengths spanning at least six orders of magnitude. Directed evolution to generate substrate specificity could use ecotin mutants iteratively, such that the inhibitory strength was increased after each round of selection.

**Figure 5.2** Theory behind the Paper, Rocks, Scissors genetic screen. A. Initially, the wild-type protease is blocked from hydrolyzing a substrate that is included in the solid growth medium and no halo is evident. B. Random mutagenesis of the protease gene will lead to mutant proteases that escape inhibition and a halo surrounding growing bacterial colonies will signal potential clones. C. Several possibilities are accounted for by this screen include the removal of non-active mutants and discrimination from enzymes with wild-type characteristics.
**Figure 5.3** Dimeric structure of ecotin. The primary binding site interacts with the target protease in a conformation identical to that observed with substrates. A secondary protease binding site is provided by the other chain.

<table>
<thead>
<tr>
<th>Protease</th>
<th>$K_i$ (nM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bovine trypsin</td>
<td>&lt;1</td>
</tr>
<tr>
<td>Factor Xa</td>
<td>54</td>
</tr>
<tr>
<td>Human leukocyte elastase</td>
<td>55</td>
</tr>
<tr>
<td>Human FXIIa</td>
<td>89</td>
</tr>
<tr>
<td>Human Kallikrein</td>
<td>163</td>
</tr>
</tbody>
</table>

**Table 5.3** Inhibition constants of ecotin against a variety of S1 family peptidases. Thrombin, activated protein C, tissue-type plasminogen activator and plasmin are poorly inhibited by wild-type ecotin.
Ecotin binds a protease in a similar fashion as a substrate and this property could be used to further control the directed evolution of protease specificity. Mutagenesis of the three residues preceding the P1 residue could be used to alter the $K_i$ of the inhibitor. The mutated sequence of the protein would be the opposite of the recognition site of the protease. For example, monomeric ecotin presenting Gln-Lys-Trp-Met in P4 to P1 should poorly inhibit coagulation factor Xa which prefers the sequence Ile-Glu-Gly-Arg. However, this mutant inhibitor should still restrict the activity of SGT. Ecotin derives from a bacterial source and production of the recombinant inhibitor in $B.\ subtilis$ should be straightforward. Given the high level of recombinant protein expression in $E.\ coli$, yields in $B.\ subtilis$ should provide a sufficient molar excess of inhibitor.

Detection of a highly specific protease in high throughput screening requires a substrate that is well defined, readily quantified, and preferably inexpensive. A commonly used method for detection of protease activity in bacterial colonies is the addition of skim milk powder to the solid medium. Protease activity leads to a zone of clearance, or halo, around the colony. Unfortunately, high specificity proteases do not hydrolyze skim milk effectively. Alternatively, peptide substrates similar to that employed in the present work can be synthesized with fluorescent leaving groups. However, these peptides are costly to produce. Green fluorescent protein (GFP) is commonly used as reporter protein for gene expression and cellular localization and could be engineered for detection of protease activity [212-214]. The fluorophore is generated from three adjacent residues, Ser-Tyr-Gly, that are sequestered on the inside an 11-stranded $\beta$-barrel (Figure 5.4) [215]. The protein displays high thermostability and extreme resistance to proteolysis resulting from the tightly packed structure [216]. Recently, Williams described three regions in GFP that could be rendered
sensitive to site specific proteolysis [217]. However, cleavage of the protein at any one of these sites did not lead to a decrease in fluorescence. These results were likely due to the stability of the β-barrel structure which did not unfold after hydrolysis of a single bond. Combining two insertion mutations lead to a protein that did not fold properly, forming inclusion bodies in their *E. coli* expression system, and further research was not continued (M. Williams, personal communication). Random mutagenesis and screening of a double insertion mutant of GFP could be readily performed to select for soluble protein that fluoresces under UV light [218]. GFP would then be a converted to a useful reagent for the detection of proteolytic activity *in vivo*. Active protease would be evident by the formation of a halo surrounding a *B. subtilis* colony if the fluorescent protein was added directly to the solid growth medium in the proposed screen.

![Figure 5.4](image)

**Figure 5.4** Structure of GFP and potential regions for insertion of a protease recognition sequence (PDB code 1EMA). For example, Ile-Glu-Gly-Arg-Ser inserted at positions 172 and 189 would allow FXa to cleave GFP twice, releasing a strand of the β-barrel, and removing the fluorescent properties of the molecule.
In summary, a directed evolution strategy for the development of novel proteolytic substrate specificities is possible. Two proteins must be created and characterized prior to validating such an approach. The recombinant expression of SGT in *B. subtilis* provides a useful starting point towards this goal and is a substantial improvement over previously reported systems [59,98]. Mutagenesis of the cleavage sequence presented by GFP and the inhibitor will permit screening for any specificity desired. Novel substrate binding pockets may result that differ from that observed elsewhere in the family, yet produce similar and more stringent substrate specificities.

### 5.5 Future Opportunities

A number of novel features exist in the trypsin scaffold that are poorly understood and could be generated on a simplified scaffold such as SGT. The role of sodium binding in generating catalytic efficiency was previously discussed as a potential avenue for research. Induced fit mechanisms of substrate specificity have been described, particularly in the complement system [219]. Autolytic activation induced by receptor binding is a well known phenomenon in the vertebrate blood coagulation cascade [220]. Zymogen activation mechanisms are known to differ in the S1 family peptidases [221-224]. Linkage of the protease domain to other protein domains could be studied to generate novel function and proteolytic activities. Lastly, little is understood in the mechanism of catalytic rate enhancement caused by phospholipid and co-factor binding in the coagulation cascade [225]. These molecular properties are likely not distinct mechanisms acting alone. Structural proximity and direct interactions with adjacent amino acid side chains suggest complex relationships with substrate binding and the catalytic process have yet to be fully understood.
5.6 Significance of the Work

In the present dissertation, a novel expression system for trypsin-like enzymes has been produced and optimized. Previous studies have for the most part used eukaryotic proteases which are inherently more difficult to work with based on their evolution to meet physiological function. I have demonstrated that primitive trypsin-like enzymes derived from a bacterial source can be used as a scaffold for engineering substrate specificity and functional properties similar to eukaryotic proteases. A number of mutations were created in SGT that increased the primary specificity for Arg containing substrates and the extended substrate specificity was improved to partially mimic FXa. These results show that the substrate specificity of any protease in the S1 family of peptidases could be duplicated using a similar approach.

Perfect mimicry of FXa substrate specificity was not achieved likely due to the requirement of additional second shell residues involved in optimization of the binding pocket. Future work could focus on further reproduction of FXa specificity through similar amino acid substitutions. An alternative approach that involves random mutagenesis and a novel genetic screen has also been described that may yield highly specific proteases that bear little sequence similarity with known proteases. Directed evolution of extended substrate specificity is technically challenging and, if successful, can significantly expand the repertoire of protease technology.

Highly selective serine proteases would be useful in a number of applications. The YSFMP mutant of SGT bearing six mutations could be used for the site specific proteolysis of recombinant proteins as an alternative to FXa. In the future, it is anticipated that proteases could be designed with levels of substrate specificity approaching that found in restriction
endonucleases [226]. These enzymes could be used for hydrolysis of proteins without the need for manipulation of the DNA sequence for addition of a protease recognition site. Moreover, peptide ligation through reverse proteolysis has been described and could be combined with highly selective mutants of SGT [227]. Thus, novel proteins could be produced more rapidly in an approach comparable to combinatorial chemistry.

5.7 Conclusions

Substrate specificity of the S1 family peptidases is derived from a few amino acids in the protein sequence. Due to the requirement of second shell residues for optimization of the substrate binding site, structure based design of selectivity is a difficult yet achievable goal. Engineering specificity is then limited by what exists in nature. Molecular evolution has not explored many of the possibilities of substrate specificity due to the physiological requirement for efficient catalysis. Directed evolution, as proposed in the present dissertation or by other means, is the next likely step in the progression of protease technology.
Appendix A: Structural Alignment of Selected SI Family Peptidases

Structures of SI family peptidases were aligned using the combinatorial extension method available on-line at http://cl.sdsc.edu/ce.html [228]. Conserved residues in all 8 polypeptides are denoted with * and mutations constructed in SGT are listed below the alignment.

---60-loop---

1SGT: 1 VGGTRAQGEFPFMVRLSM-----GCGGALYAQDIVULTAAHCV-----SGSGNTS--SIATG
1TLD: 1 IVGGTYGACTVQPSVLSNGLS----FCGGSLISSQWVSAAHCY-------K--S--GIQVRLG
3TGI: 1 IVGGTYGCAVNPSVPSVLSNGLS----FCGGSLISSQWVSAAHCY-------K--S--RIQVRLG
1PFX: 1 IVGGVENAKPQPFWVILNGLKIDA-----FCGGSLISQEFVVTAAHCPG----G------V--KITVAG
1AUT: 1 LIDGKTRGDSQPWVLPDQFELPHSAHSKCM------ACGAVLHPSWVTAAHCMDES------KLVRLG
1HCG: 1 IVQQGQEKDCPQWALLINENEG--FCGGTITLSEFYLTAAHCLAQA------RFKVVRG
1CVN: 1 IVGKVGCPKGCQVVQLNLNGQAC--LGGTTLINTLVSAAHCFKDI------K--MNRLIAVLG
1PPB: 1 IVEGSDAEIGMSPQVQLMFKSPQELCQGALSISDRVLTHALAHLPPKDF------TENDLLVRIG

---99-loop---

1SGT: 54 VVDLQ--S-G-AVKVRSTKVLQPQQYN--G-TGDKWALIKLQAPGYN------QPTLKIAT--T------T
1TLD: 52 EDINVVE-G-NEOPEISASKSTVHPNSN--SNNT-LANDMILITKSAASLNSRSVNASILPT------C
3TGI: 52 EHINVLW-G-NEOFQNAIITKHPKPF-RKL--LANDMLIKLSQVPLNARATVALPS------S--C
1PFX: 55 EYNTEETEP--TEQRRNVIRAIPHHSYNATVNKYSHDIALLELDEPLTINVFPICIAD--KEYNT---F
1AUT: 56 EYDLRRWEG--WELDLDIKEVFVHPNYS-KST--TDLNDIALLHLAQPATLSTCVICPLD-GLARREL
1HCG: 56 DRNTEQE-G-GEAVHEVVKHRRFRT-KET--YDPSIALRLKTPTITFMNQVAAPCCLRDWABS
1CVN: 58 EHDLSE-H-DGEQSRKVQVPISTVPG--TNHDIALLLRRQVPLVLCLPERFSERL--L
1PPB: 65 ESRHRHI-RNIEKLSMEIKYTHPRN-WRENLDIAMKLMKPAVFASLHVLPLCD-PRTAAS-L

96-99-172-loop

1SGT: 104 AYNQGTFTVAGWGA-----SQRYYLKKANVFPSVSDAACRSAY---GNELVANEICAGY
1TLD: 110 ASAGTCLISGWN--TKSAGT--SYDPVLKCLKAPILSDSSCSKAY------PQQT-INSMFCAG
3TGI: 110 APAGTCLISGWN--TLSCGV--NEPDLICLQADPALLPGQDAC--EAGT-DMNVCYGF
1PFX: 120 LK-FGSWSWSGWR--FVRNFR--SARATYQLYKLYPVLVDRTACLAST--KFTIY-SMNCAGF
1AUT: 120 NQAQQETLVQWG--HSSSSEKAK--NRTFVLNIKVIHYHPHCESVN------SNMVSESNLMCAGI
1HCG: 120 MT-QKTIGSVSGF--THEKGRQS-- flaMLQVPWVDNRSCKLS------SFIIT-QMNCAGF
1CVN: 122 AFVRPSLVSQGQ--LLDRG------ATAELMVVLNPRLQMDCQKVQGYDPVINNIT-BMNCAG
1PPB: 130 QAOGYKQVRGWGN--LKTETWTANVFQGPSVLSQVNLIEPVDCKST------RIRIT-DMNCAG

---172-loop---

1SGT: 160 -PDT--GGVDTCG--QDSSGGMFPRK------DNADEWIEQGVISWYGCARPGYGPTYEVSFTPASS
1TLD: 166 -L-E--GGKSDC--QDSSGGMFPRK------GLKQISWYGCARPGYGPTYEVSFTPASS
3TGI: 166 -L-E--GGKSDC--QDSSGGMFPRK------GLQISWYGCARPGYGPTYEVSFTPASS
1PFX: 174 -H-E--GGKSDC--QDSSGGMFPRK------SFLTIGISWEGECAVKGYTMYKSVRYWNI
1AUT: 180 -L-G--DRQDGAGSASPHGTH------WPLVGILWQGECGGLLHYYKVGGTYYKRVSL
1HCG: 174 -D-T--QKEDAC-EGDSSGGMFPRK------YFTVGISSWEGSVCAVDCFYTGKGSVKYHFKLM
1CVN: 181 -S-D--GKSDC--KBGSOGPHATHYRGT------WYLTVGSSWWQQCATVHGFPVTVSVQYIEL
1PPB: 191 -K-PDEAKRQVDAC---EDSSGGMFPRK------WYQMGIVSVWQGDGCDGKGYTPHYVPK

*p* ***E*** ***** M **

---190 loop---

1SGT: 217 ASAARTL Streptomyces griseus trypsin 100.0 [76]
1TLD: 217 KQTTASN Bovine beta-trypsin 35.1 [112]
3TGI: 217 QDIAAN Rat anionic trypsin 32.7 [230]
1PFX: 229 KETK-- Human coagulation factor IXa 34.0 [141]
1AUT: 235 HGLHDK Human activated protein C 32.2 [231]
1HCG: 239 DRSMKTR Human coagulation factor Xa 32.5 [157]
1CVN: 236 QKLMBSE Human coagulation factor VIIa 35.9 [158]
1PPB: 251 QKVIDOF Human alpha-thrombin 33.2 [157]
Bibliography


163. Schmidt, A. E., Padmanabhan, K., Underwood, M. C., Bode, W., Mather, T. & Bajaj, S. P. (2002). Thermodynamic linkage between the S1 site, the Na+ site, and the Ca2+ site in the protease domain of human activated protein C (APC). Sodium ion in the APC crystal structure is coordinated to four carbonyl groups from two separate loops. *J. Biol. Chem.* 277, 28987-95.


