



PROTEIN CRYSTALLOGENESIS OF THE ARROWHEAD PROTEASE INHIBITOR-A

Mémoire

LE YI LIN

Maîtrise en biochimie
Maître ès sciences (M.Sc.)

Québec, Canada

© Le Yi Lin, 2015

Résumé

Les études sur la structure atomique des macromolécules biologiques ont été d'une grande importance en révélant des relations structurales et fonctionnelles d'enzymes, d'acides nucléiques et d'autres macromolécules dans divers systèmes biologiques. Dans la présente étude, la cristallogénèse et la cristallographie aux rayons X ont été utilisés pour déterminer la structure d'inhibiteur de protéinase A à une résolution atomique.

Les inhibiteurs de protéinase A et B (API-A,-B) de la arrowhead, sont les deux principaux composés inhibiteurs qui sont purifiés à partir de la arrowhead (*Sagittaria sagittifolia*, Linn.). Ils sont des inhibiteurs à double tête multifonctionnels de diverses protéases. La structure du complexe ternaire de l'inhibiteur API lié à deux trypsines a été déterminée. Cependant, la structure tridimensionnelle de l'apoenzyme API-A est encore inconnue. A cet effet, la cristallogénèse de l'apoenzyme API-A a été réalisée. Après des étapes de purification et de cristallisation appropriées, des cristaux de l'apoenzyme API-A d'une qualité permettant la diffraction des rayons-X à haute résolution ont été obtenus en utilisant la méthode de diffusion de vapeur dans une goutte suspendue. Un ensemble complet de données de diffraction des rayons-X jusqu'à une résolution de 3,10 Å a été obtenu avec un crystal en forme de diamant, dont, le groupe d'espace a été déterminé comme étant P1. Ces données seront importantes pour comprendre la fonction d'API-A dans des systèmes biologiques.

Abstract

Structural studies of biological macromolecules at atomic resolution have revealed many specific aspects of the structure/function relationships of the enzymes, nucleic acids and other macromolecules in biological systems. In the present study, protein crystallogenesis and X-ray crystallography were used to determine the structure of the protease inhibitor A at atomic resolution.

The arrowhead protease inhibitors A and B (API-A, -B), are the two major inhibitor components which are purified from the tubes of arrowhead (*Sagittaria sagittifolia*, Linn.). Both API-A and API-B are double-headed multifunctional protease inhibitors. The ternary structure of the inhibitor API-A complex with two trypsins has been reported. However, the three-dimensional structure of the apoenzyme API-A is still unknown. In this regard, crystallogenesis of apoenzyme API-A was studied. After proper purification and crystallization steps, diffraction-quality crystals of the apoenzyme API-A were obtained, using the hanging drop vapor diffusion method. Moreover, a complete set of X-ray diffraction data was collected up to a resolution of 3.10 Å, using a diamond-shaped crystal whose space group was determined to be P1. These data will be important for understanding the function of API-A in biological systems.

Table of Contents

Résumé.....	iii
Abstract	v
Table of Contents.....	vii
List of figures	ix
List of tables	xi
List of abbreviations.....	xiii
Acknowledgements	xv
Chapter 1 Introduction	1
1.1 Protease.....	2
1.2 Serine protease	5
1.3 Serine protease inhibitor	10
1.4 Arrowhead protease inhibitor A	17
1.5 Problematics	22
1.6 Research objective	23
Chapter 2 Materials and methods.....	25
2.1 Methods in protein preparation	26
2.1.1 Cell culture.....	26
2.1.2 Protein purification	27
2.1.3 Measurement of protein concentration	29
2.1.4 Mass spectrometer	29
2.2 Methods used for protein crystallization	32
2.2.1 Preparation of protein solutions before crystallization	34
2.2.2 Vapor diffusion methods	34
2.2.3 Other methods for crystallization	36
2.2.4 New ideas in crystallization.....	39
2.2.5 Preparation of protein crystals before structure determination.....	46
2.3 Methods in protein structure determination.....	47
2.3.1 Nuclear magnetic resonance and X-ray crystallography	47
2.3.2 X-ray data collection.....	50
Chapter 3 Results and discussion.....	53
3.1 Practice in crystallization methods	54
3.1.1 Phase diagram study of trypsin: nucleation curves at various temperatures.....	54

3.1.2	The effect of composition modification.....	56
3.2	Crystallogenesi s study of API-A	59
3.2.1	Expression of API-A.....	59
3.2.2	Purification of API-A.....	62
3.2.3	Crystallization of API-A	65
3.3	X-ray data collection	67
Chapter 4	Conclusion	71
Chapter 5	Perspectives	75
Reference	77
Annex: Full data of the pET28a-derived target gene sequencing	81

List of figures

Figure 1. Schematic representation of a protein substrate binding to a protease.	2
Figure 2. Schematic representation of two possible enzyme-substrate complexes of papain with a hexapeptide.	3
Figure 3. The X-ray crystal structure of the archetypal serine protease chymotrypsin.	7
Figure 4. The catalytic triad in chymotrypsin complexes.	8
Figure 5. The generally accepted mechanism for serine proteases.	10
Figure 6. Possible mechanisms of proteinase inhibition.	16
Figure 7. Overall structure of API-A and its complex.	19
Figure 8. The views of the two interfaces.	21
Figure 9. Sequence display of API-A, from Protein Data Bank.	27
Figure 10. Mass spectrometry protocol.	31
Figure 11. Equilibration pathways in various protein crystallization methods.	33
Figure 12. Schematic representation of hanging drop, sitting drop and sandwich drop vapor diffusion method.	35
Figure 13. Classification scheme for the results of a crystallization screen.	38
Figure 14. Two-dimensional-phase diagrams of protein concentration versus PEG concentration.	40
Figure 15. Drop volume ratio effect on supersaturation.	42
Figure 16. Two-dimensional-phase diagrams of thaumatin (A) and proteinase K (B), and prediction of trajectories of varying composition modifications at 300 K and 295 K, respectively.	44
Figure 17. Process of protein structure determination by NMR (A) and X-ray crystallography (B).	49
Figure 18. Phase diagram of nucleation curves of trypsin, under different crystallization temperatures.	55

Figure 19. Agarose gel electrophoresis of DNA extracted from cultured cells.....	59
Figure 20. SDS-PAGE image of bacterial supernatant and Ni-NTA column fractions.....	62
Figure 21. SDS-PAGE image of size exclusion chromatography fractions.....	63
Figure 22. SDS-PAGE image of blue-sepharose affinity chromatography.....	64
Figure 23. Peptide analysis of the two bands API-A (above is upper band and below is lower band) by the method of mass spectrometry.....	65
Figure 24. Main shapes of crystals obtained.....	66
Figure 25. Diffraction pattern of API-A crystal.....	67

List of tables

Table 1. Families of peptidase inhibitors.....	13
Table 2. “Hits” and “Hits Increase” obtained by conventional screening and composition modification.	45
Table 3. The summary table of released entries from Protein Data Bank (PDB).....	50
Table 4. The effect of composition modification to the crystallization of glucose isomerase.....	57
Table 5. The effect of composition modification to the crystallization of β -amylase.	58
Table 6. Result of the pET28a-derived target gene sequencing.	61
Table 7. The summary data collection.....	68
Table 8. The preliminary structural data from API-A crystal.	68

List of abbreviations

Å: angstrom
Aa: amino acids
API: Arrowhead Protease Inhibitors
bp: base pair
BSA: bovine serum albumin
°C: Degree Celsius
ESI: electrospray ionization
g: gram
HPLC: high-performance liquid chromatography
IPTG: isopropyl- β -D-thiogalactopyranoside
kDa: kilodalton
L: liter
M: molar
MALDI: matrix-assisted laser desorption/ionization
 β -Me: β -mecaptoethanol
mg: milligram
min: minute
ml: milliliter
mM: millimolar
MPD: 2-Methyl-2, 4-Pentanediol
MS: mass spectrometry
MS/MS: tandem mass spectrometry
m/z: mass-to-charge ratio
ng: Nanogram
O.D.: optical density
PAGE: polyacrylamide gel electrophoresis
PCG: Protein Crystal Growth
PDB: Protein Data Bank
PEG: polyethylene glycol
PIs: protease inhibitors
PMSF: phenylmethanesulfonyl fluoride
rpm: revolutions per minute
SDS: sodium dodecyl sulfate
sec: second
serpins: serine protease inhibitors
Tris: tris(hydroxymethyl)aminomethane
 μ g: microgram
 μ l: microliter

μM : micromolar

Acknowledgements

To undertake and finish this master degree is not easy, fortunately, with the help from my supervisors, my family members, the committee members, the dean and secretary of department, I got more than that I could imagine, the master degree in biochemistry.

My supervisors professor Jacques Lapointe and professor Sheng-Xiang Lin, thank you for giving me the experience of being your student, your gentleness and patience will be remembered through all my life. During more than two years of study, I met too many problems both in study and in daily life, your kindness and patience helped me to get over those difficulties, and helped me to succeed.

I would thank my wife Zhi Qian Zhou, my parents and parents in law, they supported me all the time, gave me courage to conquer the difficulties.

My committee members, Dr. Manon Couture and Dr. Patrick Lagüe, thank you for your acknowledgements in my exams, with your great encouragement, I took back my confidence. And especially thank you for your approvals in my decision, I appreciate that a lot.

Dr. Louise Brisson, Dr. Lisa Topolnik and Mme. Véronique Bédard, I had so many problems to ask for your favor, and you all helped me dealing with them, then gave me the best results, thank you for your understanding as well as your great help.

Great thanks to Dr. Rong Shi, for offering me help in editing my memory and for those useful suggestions regarding my research.

My colleagues in CHUL, Dr. Dao-Wei Zhu, Dr. Ming Zhou, Bo Zhang, Chenyan Zhang, Xiao-Qiang Wang, Dan Xu, Jean-François Thériault, Mouna Zerradi, also very excellent colleagues in Laval University, Jonathan Huot, Sébastien Blais, Van-Hau Pham, Guillaume Bonnaure and Imane Fiher. Thank you all for your help during my study and experiments. Thanks to professor Stéphane Gagné, for his kindness on offering me the course on protein structure based on my special situation.

Chapter 1

Introduction

1.1 Protease

A protease (also termed peptidase or proteinase) is defined as an enzyme that performs proteolysis. Proteases can be found in animals, plants, bacteria, archaea and viruses. Sequencing of the human genome revealed that more than 2% of our genes encode proteases (1-5).

The protein catabolism by a protease results from the hydrolysis of the peptide bonds that link amino acids together in polypeptide chain (Figure 1). Most proteases are relatively nonspecific for their substrates, and target multiple substrates in an indiscriminate manner, whereas some proteases exhibit an exquisite specificity toward a unique peptide bond of a single protein (1).

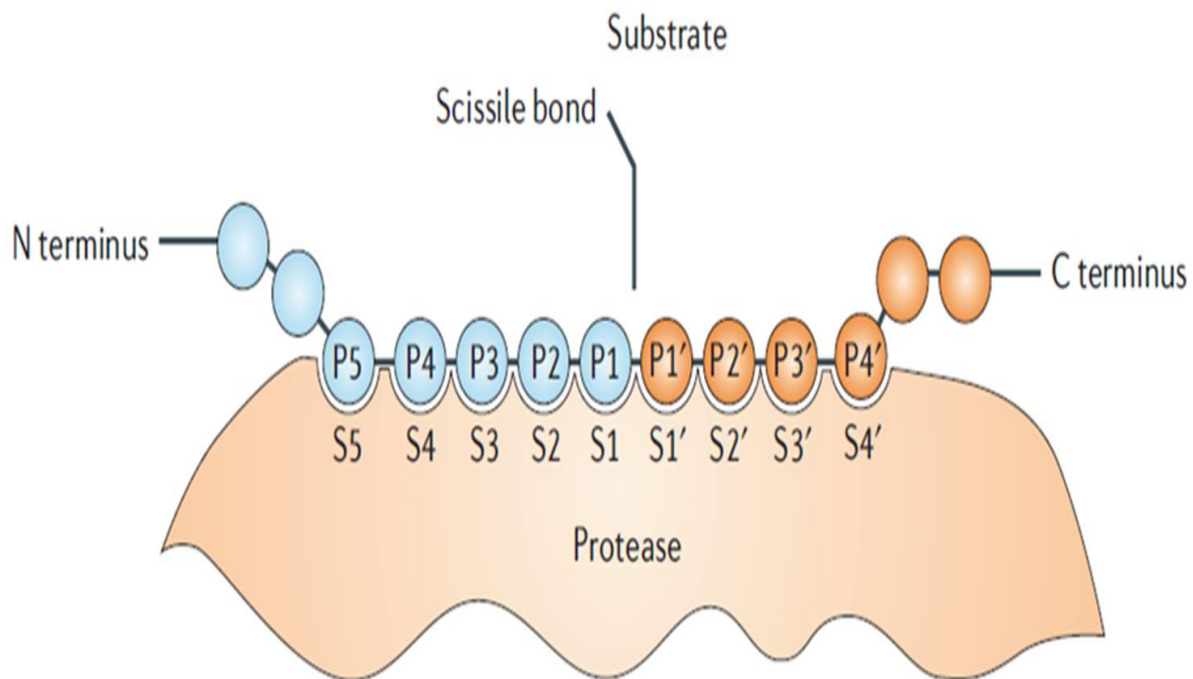


Figure 1. Schematic representation of a protein substrate binding to a protease. The surface of the protease that is able to accommodate a single side chain of a substrate residue is called the subsite. Subsites are numbered S1-Sn upwards towards the N terminus of the substrate, and S1'-Sn' towards the C terminus, beginning from the sites on each side of the scissile bond. The substrate residues they accommodate are numbered P1-Pn, and P1'-Pn', respectively (5).

Generally, the active site of an enzyme performs two functions, which are binding the substrate and catalyzing the reaction. For example, Papain is an endopeptidase (which cleaves protein substrates in the middle of the molecule) which has a large active site that extends over about 25 Å and can be divided into 7 “subsites” (Figure 2). The subsites are located on both sides of the catalytic site, 4 on the one side and 3 on the other. Each subsite accommodates one amino acid residue of the peptide substrate. The “positions” of the residues in the substrate peptide were numbered according to the “subsites” they occupy, and also depending on which bond is split. So S1’ and S1 are flanking the catalytic site, whereas hydrolysis will occur between residues P1 and P1’ (6, 7).

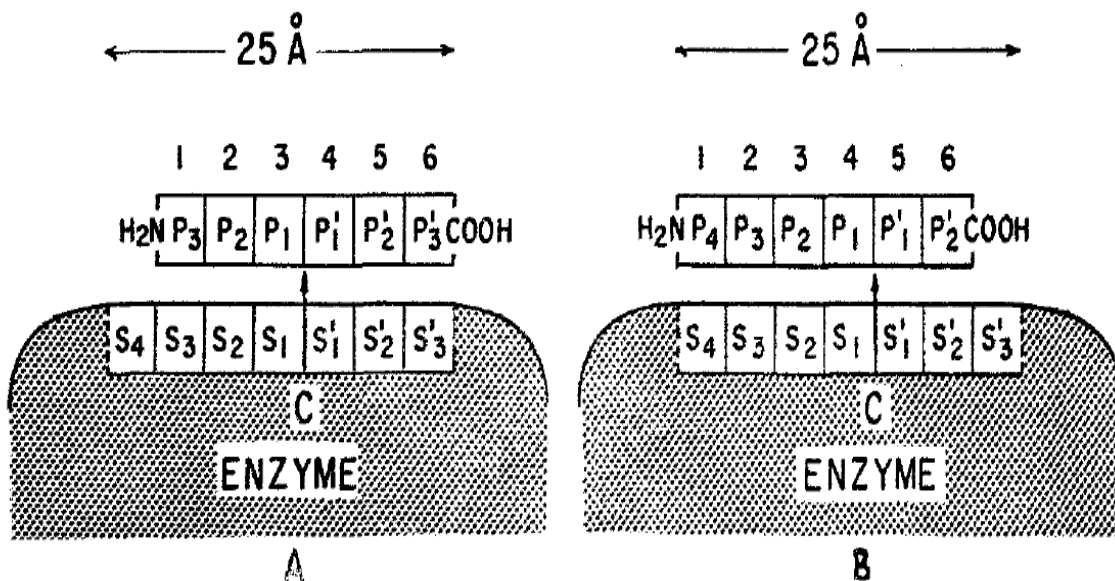


Figure 2. Schematic representation of two possible enzyme-substrate complexes of papain with a hexapeptide. The active site of the enzyme is composed of 7 ‘subsites’ (S1-S4 and S1’-S3’) located on both sides of the catalytic site C. The positions, P, on the hexapeptide substrate are counted from the point of cleavage and thus have the same numbering as the subsites they occupy. Complex A will yield as products two molecules of tripeptide, B one molecule of tetrapeptide and one of dipeptide (6).

Besides the generalized protein digestion, proteases possess complex functions. For example, proteases regulate growth factors, cytokines, chemokines and cellular receptors, both through activation and inactivation, which can lead to intracellular signaling, gene regulation, inflammation, apoptosis, blood coagulation, and embryogenesis (3-5, 8-10). Proteases are accordingly used widely in medical fields; the predominant use of proteases has been in treating cardiovascular diseases, but they are also emerging as useful agents in the treatment of sepsis, digestive disorders, inflammation, cystic fibrosis, retinal disorders, psoriasis and other diseases (4). Up-regulation of proteolysis is associated commonly with different types of cancer and is linked to tumor metastasis, invasion and growth (9). Proteases also play key roles in plants, where they contribute to the processing, maturation, or destruction of specific sets of proteins in response to developmental cues or to variations in environmental conditions (11). Likewise, many infectious microorganisms require proteases for replication or use proteases as virulence factors, which has facilitated the development of protease-targeted therapies for diseases of great relevance to human life such as AIDS (5). Proteases are also important tools of the biotechnological industry where they are used as biochemical reagents or in the manufacture of numerous products (11).

Proteases have evolved multiple times. The comparisons of the amino acid sequences, three-dimensional structures, and enzymatic reaction mechanisms of proteases, have revealed the existence of distinct families, each of them performing the same proteolytic reactions by completely different catalytic mechanisms (1, 2, 5, 8).

Proteases were initially classified into endopeptidases, which cleave protein substrates in the middle of the molecule, and exopeptidases, which cleave protein from the N or C termini (aminopeptidases and carboxypeptidases, respectively), the action of which is directed by the NH₂ and COOH termini of their corresponding substrates. However, new classification schemes were made according to the availability of structural and mechanistic information on these enzymes. Based on the mechanism of catalysis, proteases are classified into six distinct classes:

aspartic, glutamic, and metalloproteases, cysteine, serine, and threonine proteases. Glutamic proteases have not been found in mammals so far. The first three classes, aspartic, glutamic, and metalloproteases, utilize an activated water molecule as a nucleophile, to attack the peptide bond of the substrate, whereas for cysteine, serine, and threonine proteases, the nucleophile is an amino acid residue (Cys, Ser, or Thr, respectively) located in the active site. By amino acid sequence comparison, proteases of the different classes can be further grouped into families, and families can be assembled into clans based on similarities in their three-dimensional structures. Metalloproteases and serine proteases are the most densely populated classes, with 194 and 176 members, respectively, followed by 150 cysteine proteases, whereas threonine and aspartic proteases contain only 28 and 21 members, respectively (1, 5).

1.2 Serine protease

The serine proteases, also called serine endopeptidases, are characterized by a uniquely reactive serine side chain. The serine proteases are widely distributed and of diverse functions, and are found in both eukaryotes and prokaryotes. Almost one-third of all proteases can be classified as serine proteases (12, 13).

Serine proteases can be classified into two broad categories based on their structure: chymotrypsin-like (trypsin-like) and subtilisin-like (14). In humans, they are responsible for co-ordinating various physiological functions, including digestion, immune response, blood coagulation and reproduction (13).

Chymotrypsin-like serine proteases are the most abundant in nature. They have a distinctive structure, which consists of two β -barrel domains that converge at the catalytic active site. According to the substrate specificity, chymotrypsin-like serine proteases can be further categorized as trypsin-like (positively charged residues Lys/Arg preferred at P1),

chymotrypsin-like (large hydrophobic residues Phe/Tyr/Leu at P1) or elastase-like (small hydrophobic residues Ala/Val at P1) (15).

A good example of this category is chymotrypsin. It has 245 residues, arranged in two six-stranded beta barrels (Figure 3). The active site cleft is located between the two barrels. This structure is generally divided into catalytic, substrate recognition and zymogen (an inactive enzyme precursor which requires a biochemical change to become an active enzyme) activation domain components. These three processes involve many of the same structural features and are intricately intertwined. Moreover, these domains are common to all chymotrypsin-like serine proteases (13).

Subtilisin, a serine protease found in prokaryotes, is evolutionarily not related to the chymotrypsin-clan, but share the same catalytic mechanism, which is using a catalytic triad to create a nucleophilic serine. The comparison of chymotrypsin and subtilisin is the classical example used to illustrate convergent evolution, since the same mechanism evolved twice independently in topologically different proteins during evolution (12, 13).

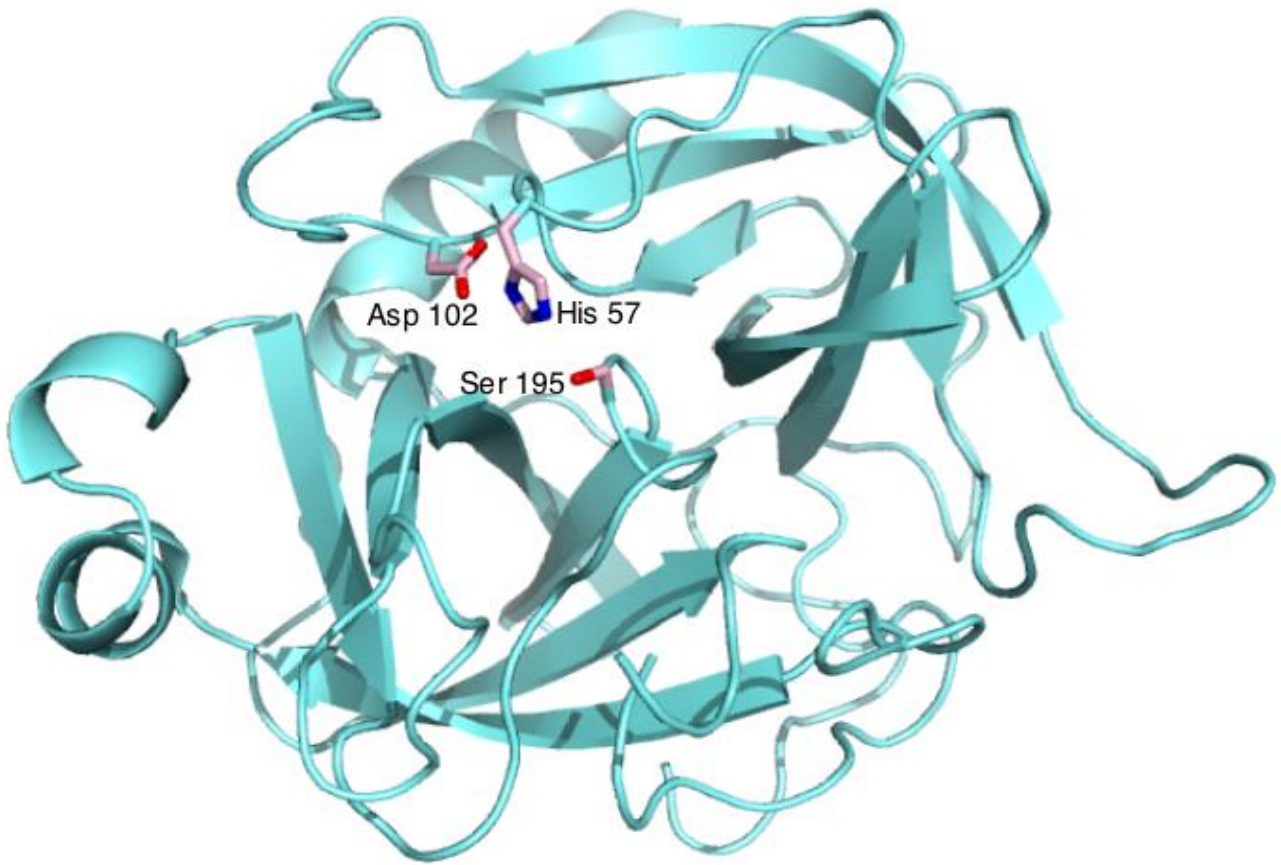


Figure 3. The X-ray crystal structure of the archetypal serine protease chymotrypsin (16). The three catalytic residues (His 57, Asp 102 and Ser 195) are labeled. PDB number: P00766 (CTRA_BOVIN).

The main function for all proteases is to hydrolyze peptide bonds. To achieve this, there are three obstacles to overcome: (a) amide bonds are very stable due to electron donation from the amide nitrogen to the carbonyl group. Proteases usually use a general acid to interact with the carbonyl oxygen, to activate an amide bond of the substrate; (b) water is a poor nucleophile; proteases always activate water, usually via a general base; and (c) amines are poor leaving groups; proteases protonate the amine prior to expulsion. These mechanisms can also be used to hydrolyze other acyl compounds, including amides, anilides, esters, and thioesters (13). Before carrying out the proteolysis reactions, the proteases must establish themselves in appropriate locations in the cellular environment. Different strategies can be used for their localization: in

most cases, they are present in complex networks, which include other proteases, substrates, cofactors, inhibitors, adaptors, receptors, and binding proteins (1).

The reaction mechanism of the serine proteases was originally revealed by the presence of the Asp-His-Ser “charge relay” system or “catalytic triad”, in which serine serves as the nucleophilic amino acid at the active site (Figure 4). More recently, serine proteases with novel catalytic triads and dyads have been discovered, including Ser-His-Glu, Ser-Lys/His, His-Ser-His, and N-terminal Ser (16, 17).

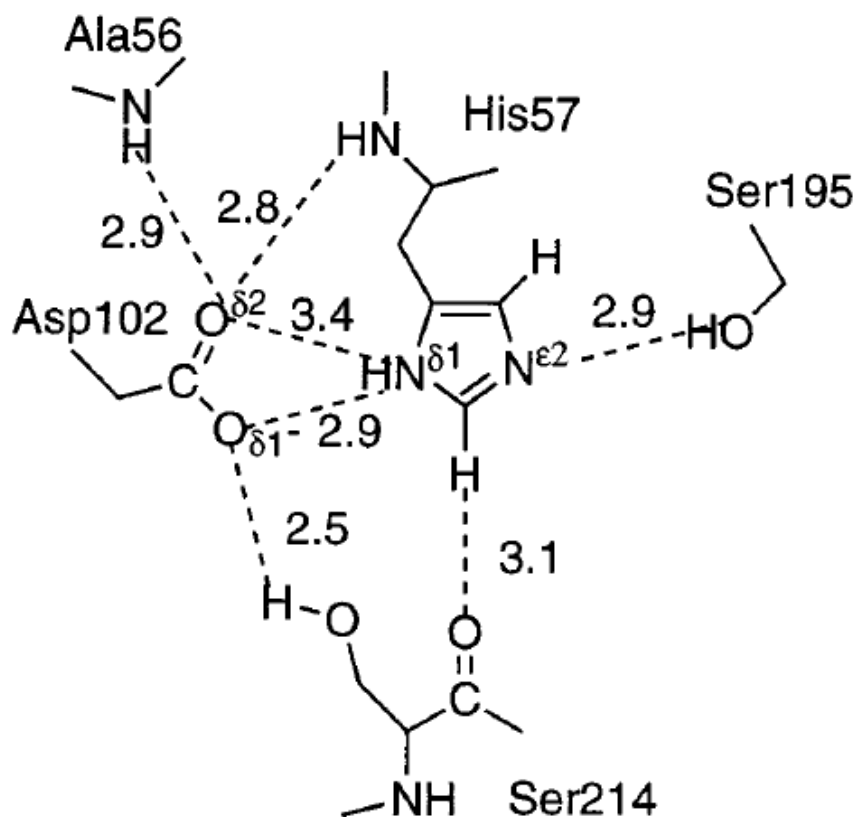


Figure 4. The catalytic triad in chymotrypsin complexes. The hydrogen bonding network of the catalytic triad is depicted for the complex with the protein inhibitor eglin C. The dotted lines represent potential hydrogen bonds (13).

The catalytic triad is part of an extensive hydrogen bonding network. It spans the active site cleft, with Ser195 on one side and Asp102 and His57 on the other (Figure 3). Generally, hydrogen bonds are shown between the N δ 1-H of His57 and O δ 1 of Asp102, also between the OH of Ser195 and the N ϵ 2-H of His57, although the latter hydrogen bond is lost when His57 is protonated. The complexes of chymotrypsin and protein inhibitor eglin C are discussed to explain the mechanism in serine protease (Figure 4). The hydrogen bonds are observed throughout the catalytic cycle. Moreover, the OH of Ser214 forms a hydrogen bond with O δ 1 of Asp102, this hydrogen bond can be found in almost all chymotrypsin-like proteases. Accordingly, Ser214 was once considered the fourth member of the catalytic triad, although more recent evidence indicates that it is dispensable. Hydrogen bonds are also observed between the O δ 2 of Asp102 and the main chain NHs of Ala56 and His57. These hydrogen bonds are believed to orient Asp102 and His57. Additionally, a novel hydrogen bond is observed between the C ϵ 1-H of His57 and the main chain carbonyl of Ser214, the carbonyl oxygen of Ser214 also plays a part in the polypeptide binding site, so it was considered to mediate the communication between substrate and the catalytic triad. Similar hydrogen bonds are also observed in subtilisin and other classes of serine hydrolases, which suggests that this interaction may be important for catalytic triad function (13, 18).

An oxyanion hole (a pocket in the structure of an enzyme which stabilizes a deprotonated oxygen or alkoxide, often by placing it close to positively charged residues) was believed to exist and it is formed by the backbone NHs of Gly193 and Ser195 (Figure 5). These atoms form a pocket of positive charge, and the pocket can activate the carbonyl of the scissile peptide bond and stabilize the negatively charged oxyanion of the tetrahedral intermediate. The oxyanion hole is structurally linked to the catalytic triad and the Ile16-Asp194 salt bridge via Ser195. Subtilisin also contains an oxyanion hole, formed by the side chains of Asn155, Thr220, and the backbone NH of Ser221 (13, 19).

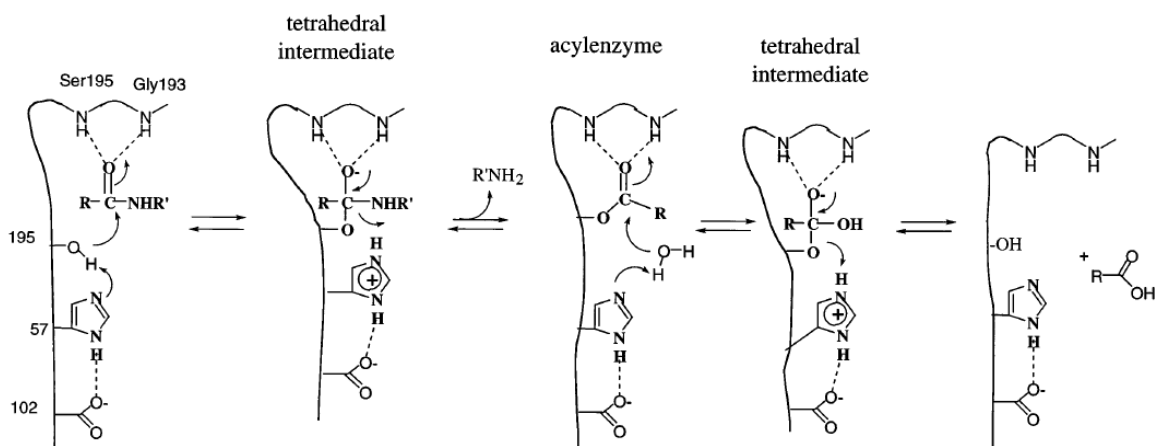


Figure 5. The generally accepted mechanism for serine proteases (13).

The generally accepted mechanism for chymotrypsin-like serine proteases is shown (Figure 5). In the acylation half of the reaction, Ser195 attacks the carbonyl of the peptide substrate, with the assistance by His57 who acts as a general base, to yield a tetrahedral intermediate. The resulting His57-H⁺ is stabilized by the hydrogen bond to Asp102. The oxyanion of the tetrahedral intermediate is stabilized by interaction with the main chain NHs of the oxyanion hole. Then the tetrahedral intermediate collapses with expulsion of the leaving group, assisted by His57-H⁺ acting as a general acid, to yield the acylenzyme intermediate. The deacylation half of the reaction repeats the above sequence: water attacks the acylenzyme, assisted by His57, yielding a second tetrahedral intermediate. Finally this intermediate collapses, expelling Ser195 and carboxylic acid product (12, 13).

1.3 Serine protease inhibitor

As they catalyze the cleavage of peptide bonds, the proteases can be potentially very damaging in living systems, so their activities need to be kept strictly under control. The action of proteases

can be controlled *in vivo* by several mechanisms: regulation of gene expression; activation of their inactive zymogens; blockade by endogenous inhibitors; targeting to specific compartments such as lysosomes, mitochondria, and specific apical membranes; and post-translational modifications such as glycosylation, metal binding, S–S bridging, proteolysis, and degradation (1). Among these mechanisms, the interactions between proteases and their inhibitors can be very efficient and important (20, 21).

Protease inhibitors (PIs) can regulate their corresponding proteases in a very significant way. They are widely distributed in all organisms, including microorganisms, plants and animals. Many of these inhibitors have a much larger size than their target proteinases. Only some microorganisms produce and secrete small non-proteinaceous inhibitors, which impair the proteolytic activity of host proteinases (21).

Protease inhibitors are widely used for clinical applications. Some inherited diseases attributable to abnormalities of proteases were proved to be susceptible to treatment with the inhibitors administered as drugs, with synthetic inhibitors that take over their function or with the natural inhibitors made available by gene therapy (20). There is great interest in developing more potent therapeutic PIs for treating human diseases related to cancer (22), pancreatitis (23), thrombosis (24), and AIDS (25).

The number of proteinaceous protease inhibitors isolated and identified so far is extremely large. These inhibitors can have very different polypeptide scaffolds, and some of them are bis- or multi-headed tandem proteins (20, 21, 26). They have been classified into families of related structures or according to the catalytic class of their target proteases. However, this classification is hampered by the occurrence of both compound inhibitors and pan-inhibitors; the compound inhibitors contain inhibitor units of different protease classes, whereas the pan-inhibitors target enzymes of different classes through a trapping reaction induced after inhibitor cleavage by the targeted protease (1, 27).

Protease inhibitors can be classified functionally as single-headed (if they have only one active reactive site), double-headed (if they have two) and so on (28). They can also be classified according to their mechanism of inhibition into four groups: the canonical inhibitors using substrate-like binding mode; exosite-binding inhibitors (exosites are the regions outside of the active site that influences catalysis), which bind a region adjacent to the active site; a third group of protease inhibitors use an intermediate mechanism based on a combination of the canonical and exosite-binding mechanisms; finally, allosteric inhibitors bind a region that is distantly located from the active site (1, 21).

In the paper of Laskowski and Kato (28), protease inhibitors could best be classified in their homologous families, but at that time the available sequence information was very limited. In recent years, a system was created, in which the inhibitor units of the protease inhibitors are assigned to 48 families on the basis of similarities detectable at the level of amino acid sequence. Then, on the basis of their three-dimensional structures, 31 of the families are assigned to 26 clans (20).

In most cases, all members of a specific inhibitor family are directed against the target proteinases of the same mechanistic class; only a very few protein inhibitors exhibit 'dual' activity simultaneously exerted towards proteinases from different mechanistic classes.

Table 1. Families of peptidase inhibitors (20).

Family or subfamily	Common name	Type-example-inhibitor-unit name (source)	SWISS-PROT accession (inhibitor-unit range); Pfam accession	Families of peptidases inhibited
I1	Kazal	ovomucoid unit 3 (<i>Meleagris gallopavo</i>)	P01004 (135–185); PF00050	S1 [15]
I2	Kunitz (animal)	aprotinin (<i>Bos taurus</i>)	P00974 (36–93); PF00014	S1 [15]
I3A	Kunitz (plant)	soybean trypsin inhibitor (<i>Glycine max</i>)	P01070 (25–205); PF00197	Mainly S1 [15], but also C1 [25–27] and A1 [28]
I3B		proteinase inhibitor B (<i>Sagittaria sagittifolia</i>)	P07479 (25–181); PF00197	S1 [15]
I4	serpin	α_1 -proteinase inhibitor (<i>Homo sapiens</i>)	P01009 (25–418); PF00079	Mainly S1 [29], but also S8 [30], C1 [31,32] and C14 [33]
I5	ascidian	ascidian trypsin inhibitor (<i>Halocynthia roretzi</i>)	P16589 (1–55)	S1 [34]
I6	cereal	ragi seed trypsin/ α -amylase inhibitor (<i>Eleusine coracana</i>)	P01087 (1–122); PF00234	S1 [35]
I7	squash	trypsin inhibitor MCTI-1 (<i>Momordica charantia</i>)	P10294 (1–30); PF00299	S1 [36]
I8	Ascaris	nematode anticoagulant inhibitor (<i>Ascaris suum</i>)	P07851 (1–63); PF01826	S1 [37], but also M4 [38]
I9	YIB	protease B inhibitor (<i>Saccharomyces cerevisiae</i>)	P01095 (1–74)	S8 [39]
I10	marinostatin	marinostatin (<i>Alteromonas</i> sp.)	P29399 (1–14)	S1 [40]
I11	ecotin	ecotin (<i>Escherichia coli</i>)	P23827 (21–162); PF03974	S1 [41]
I12	Bowman-Birk	Bowman-Birk plant trypsin inhibitor (<i>Glycine max</i>) unit 1	P01055 (42–71); PF00228	Mainly S1 [42], but also C1 [43]
I13	pot 1	eglin C (<i>Hirudo medicinalis</i>)	P01051 (1–70); PF00280	Mainly S1 [44], but also S8 [45]
I14	hirudin	hirudin (<i>Hirudo medicinalis</i>)	P01050 (1–65); PF00713	S1 [46]
I15	antistasin	antistasin unit 1 (<i>Haemonteria officinalis</i>)	P15358 (18–72); PF02822	S1 [47]
I16	SSI	subtilisin inhibitor (<i>Streptomyces albogriseolus</i>)	P01006 (32–144); PF00720	Mainly S8 [48], but also S1 [49] and M4 [50]
I17	elafin	mucus proteinase inhibitor unit 2 (<i>Homo sapiens</i>)	P03973 (26–83); PF00095	S1 [51]
I18	mustard	mustard trypsin inhibitor (<i>Sinapis alba</i>)	P26780 (31–93); PF05828	S1 [52]
I19	pacifastin	proteinase inhibitor LCMI I (<i>Locusta migratoria</i>)	P80060 (20–54)	S1 [53]
I20	pot 2	proteinase inhibitor II (<i>Solanum tuberosum</i>)	P01080 (28–86); PF02428	S1 [54]
I21	7B2	secretogranin V (<i>Homo sapiens</i>)	P05408 (27–212); PF05281	S8 [55]
I24	pinA	pinA endopeptidase La inhibitor (bacteriophage T4)	P07068 (1–161)	S16 [56]
I25A	cystatin 1	cystatin A (<i>Homo sapiens</i>)	P01040 (1–98); PF00031	C1 [57]
I25B	cystatin 2	ovocystatin (<i>Gallus gallus</i>)	P01038 (24–139); PF00031	Mainly C1 [58], but also C13 [59]
I25C	cystatin 3	metalloprotease inhibitor (<i>Bothrops jararaca</i>)	Q9DGI0 (28–141); PF00031	Not C1, but S8 [60], M12 [61]
I27	calpastatin	calpastatin unit 1 (<i>Homo sapiens</i>)	P20810 (170–222); PF00748	C2 [62]
I29	CTLA	cytotoxic T-lymphocyte antigen	P12399 (27–136)	C1 [63]
I31	thyropin	equistatin (<i>Actinia equina</i>)	P81439 (50–98); PF00086	C1 [64]
I32	IAP	BIRC-5 protein (<i>Homo sapiens</i>)	O15392 (1–142); PF00653	C14 [65]
I33	ascaris PI3	ascaris pepsin inhibitor PI-3 (<i>Ascaris suum</i>)	P19400 (21–169)	A1 [66]
I34	IA3	saccharopepsin inhibitor (<i>Saccharomyces cerevisiae</i>)	P01094 (1–68)	A1 [67]
I35	timp	timp-1 (<i>Homo sapiens</i>)	P01033 (24–207); PF00965	Mainly M10 [68], but also M12 [69]
I36	SMI	<i>Streptomyces</i> metalloproteinase inhibitor (<i>Streptomyces nigrescens</i>)	P01077 (30–131)	M4 [70]
I37	PCI	potato carboxypeptidase inhibitor (<i>Solanum tuberosum</i>)	P01075 (1–39)	M14 [46]
I38	aprin	metalloproteinase inhibitor (<i>Erwinia chrysanthemi</i>)	P18958 (20–120); PF02977	M10 [71]
I39	α_2 M	α_2 -macroglobulin (<i>Homo sapiens</i>)	P01023 (24–1474); PF00207	Numerous families including aspartic, cysteine, metallo and serine catalytic types [72]
I40	bombyx	<i>Bombyx</i> subtilisin inhibitor (<i>Bombyx mori</i>)	Q10731 (1–77)	S8 [73]
I42	chagasin	chagasin (<i>Leishmania major</i>)	Q9GY64 (1–179)	C1 [74]
I43	oprin	oprin (<i>Didelphis marsupialis</i>)	P82957 (83–291); PF00047	M12 [75]
I44	–	carboxypeptidase A inhibitor (<i>Ascaris suum</i>)	P19399 (1–65)	M14 [76]
I46	LCI	leech carboxypeptidase inhibitor (<i>Hirudo medicinalis</i>)	P81511 (16–81)	M14 [77]
I47	latexin	latexin (<i>Homo sapiens</i>)	Q9BS40 (1–222)	M14 [78]
I48	clitocypin	clitocypin (<i>Lepista nebularis</i>)	Q9P4A2 (1–150)	C1 [79]
I49	proSAAS	proSAAS (<i>Homo sapiens</i>)	Q9UHG2 (34–260)	S8 [80]
I50	p35	baculovirus p35 caspase inhibitor (<i>Spodoptera litura</i> nucleopolyhedrovirus)	O41275 (1–296); PF02331	C14 [81], but also C25 [82]
I51	IC	carboxypeptidase Y inhibitor (<i>Saccharomyces cerevisiae</i>)	P14306 (1–219); PF01161	S10 [83]
I52	TAP	tick anticoagulant peptide (<i>Omithodoros moubata</i>)	P17726 (1–60)	S1 [84]
I57	–	staphostatin B (<i>Staphylococcus aureus</i>)	Q9EYW6 (1–109)	C47 [85]
I58	–	staphostatin A (<i>Staphylococcus aureus</i>)	Q99SX7 (1–108)	C47 [85]
I59	triabin	triabin (<i>Triatoma pallidipennis</i>)	Q27049 (19–160)	S1 [86]

The ways by which the inhibitors interact with their target enzymes vary enormously. Concerning the interactions between the inhibitors and their target proteases, two general types

can be recognized: irreversible ‘trapping’ reactions, and reversible tight binding interactions (20).

The irreversible trapping reaction depends most directly on the peptidase activity of the target protease; it is covalently modified and thereby inhibited by its proteolytic activation of the protease inhibitors. This kind of inhibitor acts as a substrate, utilizes the enzymes catalytic machinery to trap and inhibit the protease. This kind of reaction is specific for endopeptidases because it depends upon the cleavage of an internal peptide bond within the inhibitor, cleavage that triggers a conformational change. If the target protease is in its catalytically inactive form, it will fail to enter into a trapping reaction, no matter whether it may well bind tightly to an inhibitor from one of the reversible classes. Trapping reactions are never truly reversible because the inhibitors are modified and reformed, which can also be described as a suicide inhibitor or suicide substrates.

For the reversible tight-binding interactions, the inhibitor makes high-affinity interactions with the active site of the target enzyme. The mechanism here is termed the ‘standard’ mechanism. In a standard mechanism inhibitor, the inhibitory unit has a single reactive-site peptide bond, and inhibition is caused by the binding of the inhibitor to the target enzyme in a substrate-like fashion.

The reversible tight binding interactions mechanisms between proteases inhibitors and their target proteases can be classified in four kinds (Figure 6). Firstly, for the canonical inhibitors, they bind the active site of their target proteases and block directly the active site in a virtually substrate-like manner. Secondly, the exosite-binding inhibitors, do not bind the active site of target proteases to directly block the catalytic residues, but they bind a region adjacent to the active site, thereby preventing substrate access to this center. The exosite binding provides two major benefits: it increases the surface area of the protein-protein interaction, leading to a greater affinity, and it can have a significant effect on the specificity of the inhibitor. The third kind of mechanism is carried out by a group of protease inhibitors using an intermediate mechanism

based on a combination of the canonical and exosite-binding mechanisms. Finally, the mechanism of allosteric inhibitors includes the binding of a region that is distantly located from the active site, but this binding prevents dimerization of the target protease and can effectively block its activity (21).

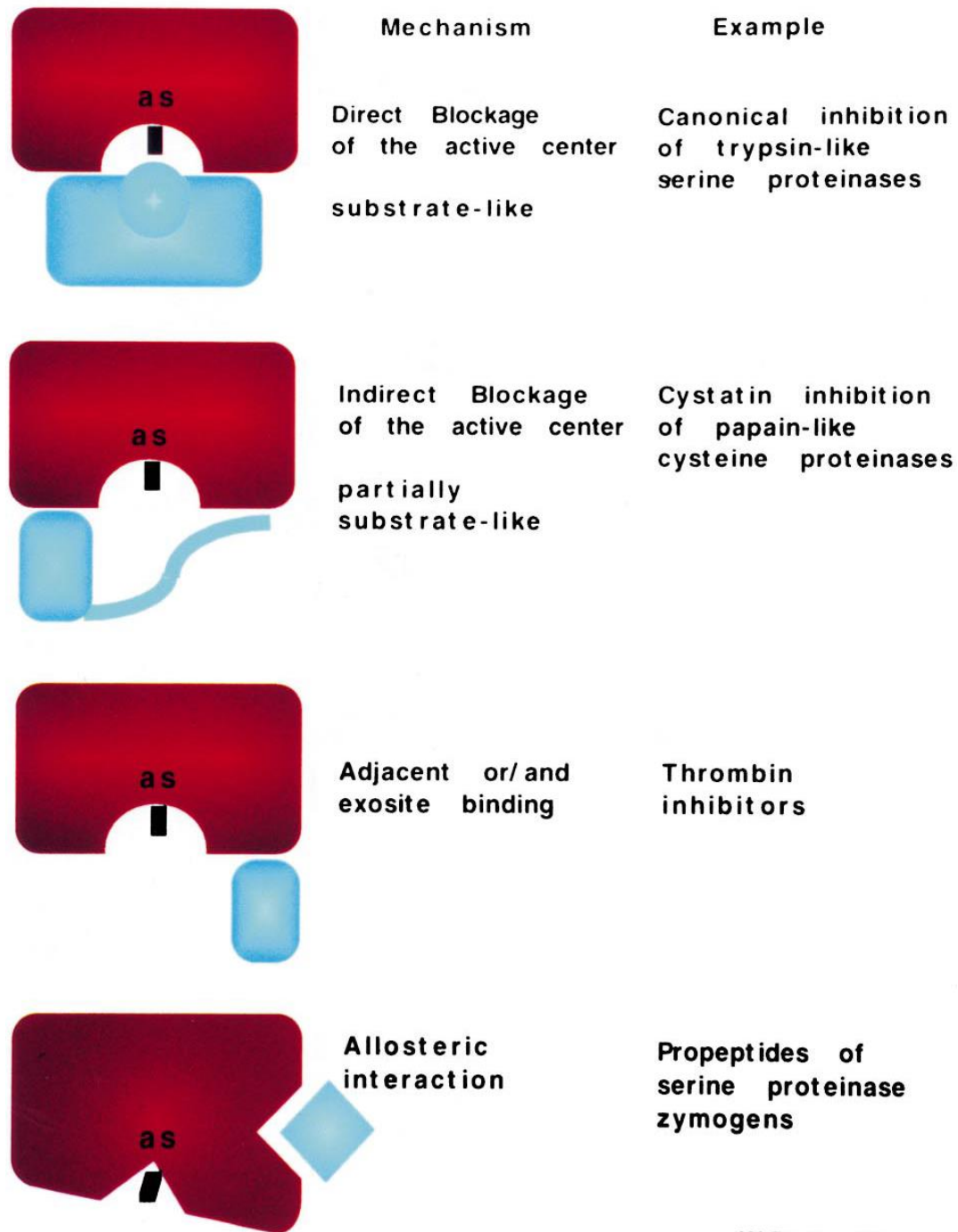


Figure 6. Possible mechanisms of proteinase inhibition (21).

1.4 Arrowhead protease inhibitor A

The inhibitors of soybean Kunitz-type serine proteases generally contain 170–200 residues and have two disulfide bonds. Most members have only one reactive site, normally located in the region of residues 60–70 (29-31). However, a few members termed double-headed inhibitors have two reactive sites, which can simultaneously bind two protease molecules; these inhibitors are classified into family I3 of peptidase inhibitors (Table 1) (20). Most members of this type are grouped into subfamily I3A. However, owing to their very low sequence similarity to other members, the double-headed arrowhead protease inhibitors API-A and API-B, purified from arrowhead *Sagittaria sagittifolia*, Linn., are grouped in subfamily I3B (20). API-A and API-B, consist of 179 residues and three disulfide bonds. They can inhibit various serine proteases, including trypsin, chymotrypsin, and porcine tissue kallikrein (32, 33). The API-A and API-B share 91% of sequence identity, though their inhibitory specificities differ. API-A can bind one molecule of trypsin and another molecule of chymotrypsin, whereas the API-B can bind two molecules of trypsin simultaneously (34).

The overall fold of API-A belongs to the β -trefoil fold with a core of six antiparallel β -strands (β 1 to β 12) surrounded by 13 loops, which resembles that of the Kunitz-type trypsin inhibitors. The hydrophobic core consists of a shallow β -barrel formed by three β -ribbons (β 1- β 12, β 4- β 5 and β 8- β 9) and a cap of three hairpins (β 2- β 3, β 6- β 7 and β 10- β 11). A 3_{10} helix is located between loop 8 and loop 9. In addition, API-A contains three disulfide bonds. The first bond (Cys⁴³–Cys⁸⁹) cross-links loop 3 and loop 6 of reactive site 1 (RS1), whereas the other two bonds are located in RS2, the one of (Cys¹⁴¹–Cys¹⁴⁴) forms an intraloop disulfide bond in loop 10, the other one of (Cys¹³⁹–Cys¹⁴⁸) stabilizes loop 10 with strand β 9. These disulfide bonds were considered to play a role in stabilizing the conformation of the reactive sites (35) (Figure 7A). The reactive sites 1 (RS1) and 2 (RS2) of API-A were identified as Leu⁸⁷ and Lys¹⁴⁵, respectively, and both reactive sites adopt a typical noncovalent “lock and key” inhibitory mechanism in a substrate-like manner.

Reactive site 1, composed of residues P5 Met⁸³ to P5' Ala⁹², has a novel conformation with the Leu⁸⁷, which is embedded in the S1 pocket completely, even though it is not a favorable P1 residue for trypsin. Reactive site 2, composed of residues P5 Cys¹⁴¹ to P5' Glu¹⁵⁰, the P1 residue Lys¹⁴⁵ adopts a classic mode binding with trypsin, using a two-disulfide-bonded loop (Figure 7, 8).

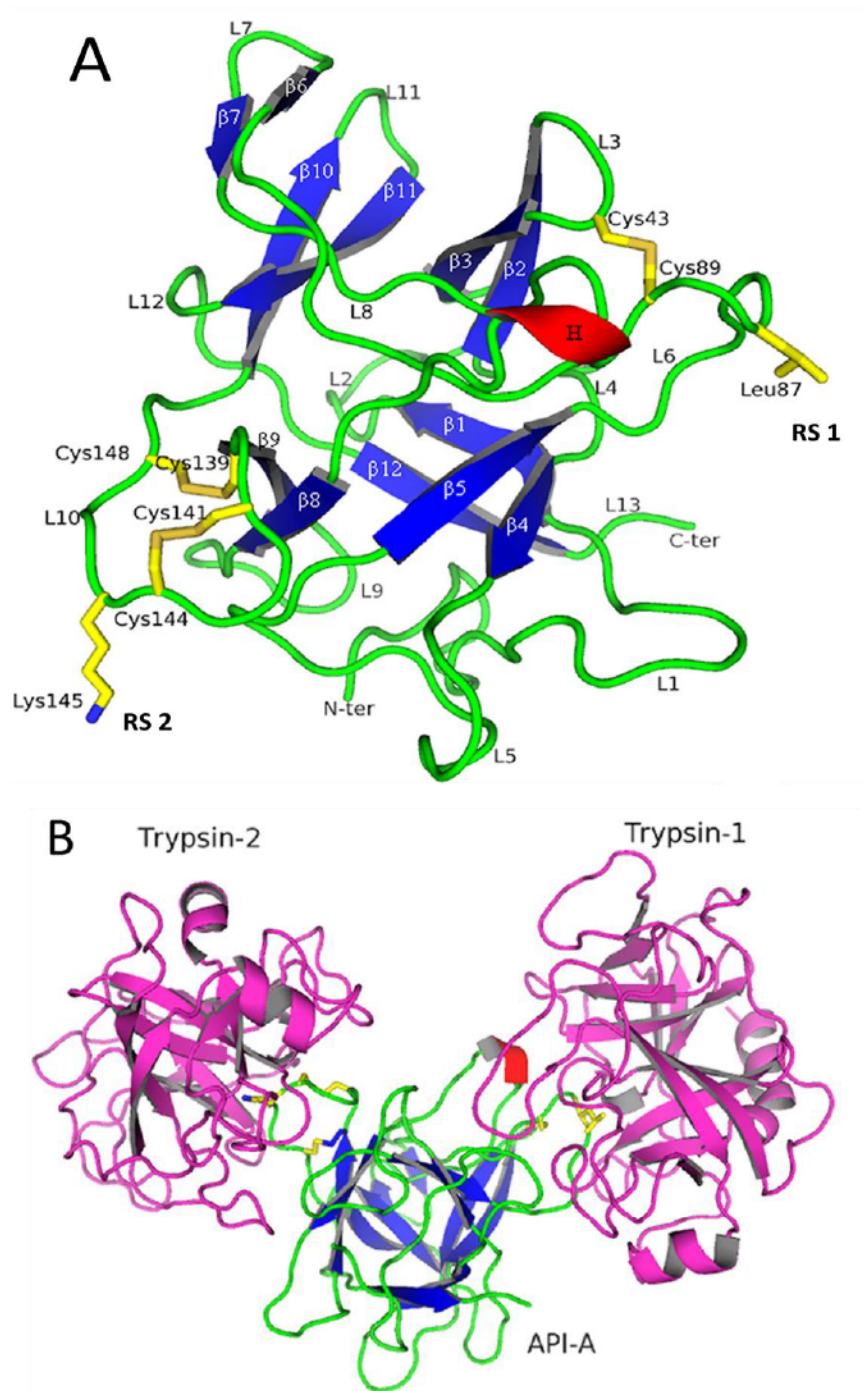


Figure 7. Overall structure of API-A and its complex. A, overall structure of API-A, with β -strands and loops numbered sequentially. Two P1 residues and six cysteine residues are highlighted as *yellow sticks*. B, overall structures of the API-A ternary complex with trypsin. API-A is colored by secondary structure assignment: 3_{10} helix in *red*, strands in *blue*, and loops in *green*. The trypsin molecules are colored *magenta* (B) (35).

The crystal structure of API-A in complex with two molecules of bovine trypsin (Figure 7B), was the first report on the three-dimensional structure of this double-headed Kunitz-type trypsin inhibitor in complex with two molecules of protease. It shows a ternary complex of the double-headed inhibitor API-A simultaneously bound to two trypsin molecules, where the two reactive sites of API-A adopt different conformations (35).

At the reactive site 1, residues P1' to P3' of API-A adopt a novel conformation, whereas a hydrogen bond network bridges loops 6 and 8 (Ile⁸⁸-O-Phe¹¹⁵-N, Asp⁹⁰-N-Asp¹¹³-O, and Asp⁹⁰-Oδ2-Asp¹¹³-N); such a conformation has not been reported for other Kunitz-type inhibitors. Loop 6 becomes more rigid because of the intraloop hydrogen bond between Pro⁸⁶-O and Cys⁸⁹-N. Moreover, the interloop disulfide bond Cys⁴³-Cys⁸⁹ between loops 6 and 3 further stabilizes the first reactive site loop and thus also contributes to the novel conformation of RS1.

The reactive site 2 adopts a canonical conformation even though the presence of two disulfide bonds (Cys¹³⁹-Cys¹⁴⁸ and Cys¹⁴¹-Cys¹⁴⁴) makes it different from other protease inhibitors that contain only one disulfide bond.

Notably, the three important spacer residues Arg⁷⁶, Glu¹²⁴, and Arg¹²⁶ adopt a unique interaction pattern that is different from the patterns in other Kunitz-type inhibitors. Arg⁷⁶ forms two direct hydrogen bonds with P2 (Arg⁷⁶-Nη2-Cys¹⁴⁴-O) and P4 (Arg⁷⁶-Nε-Glu¹⁴²-O) and one indirect hydrogen bond with P4 (Arg⁷⁶-O-Glu¹⁴²-Oε2) that is mediated by a water molecular, Wat⁶⁷. The hydrogen bonds contributed by Arg⁷⁶ are important for stabilizing RS 2, thus making it indispensable for maintaining the inhibitory activity.

The two other spacer residues Glu¹²⁴ and Arg¹²⁶ also participate in two salt bridges and one hydrogen bond (Glu¹²⁴-Oε1-Arg⁷⁶-Nη1, Glu¹²⁴-Oε2-Arg¹²⁶-Nε, and Arg¹²⁶-Nη2-Cys¹⁴⁸-O). In addition, P4 Glu¹⁴² and P4' Pro¹⁴⁹ form hydrogen bonds with the neighboring residues Ser⁶⁴ (Ser⁶⁴-N-Glu¹⁴²-Oε2) and Ala¹³⁸ (Ala¹³⁸-N-Pro¹⁴⁹-O), respectively. This interaction network in addition to two disulfide bonds gives loop 10 a well defined conformation that aids in the inhibitory activity.

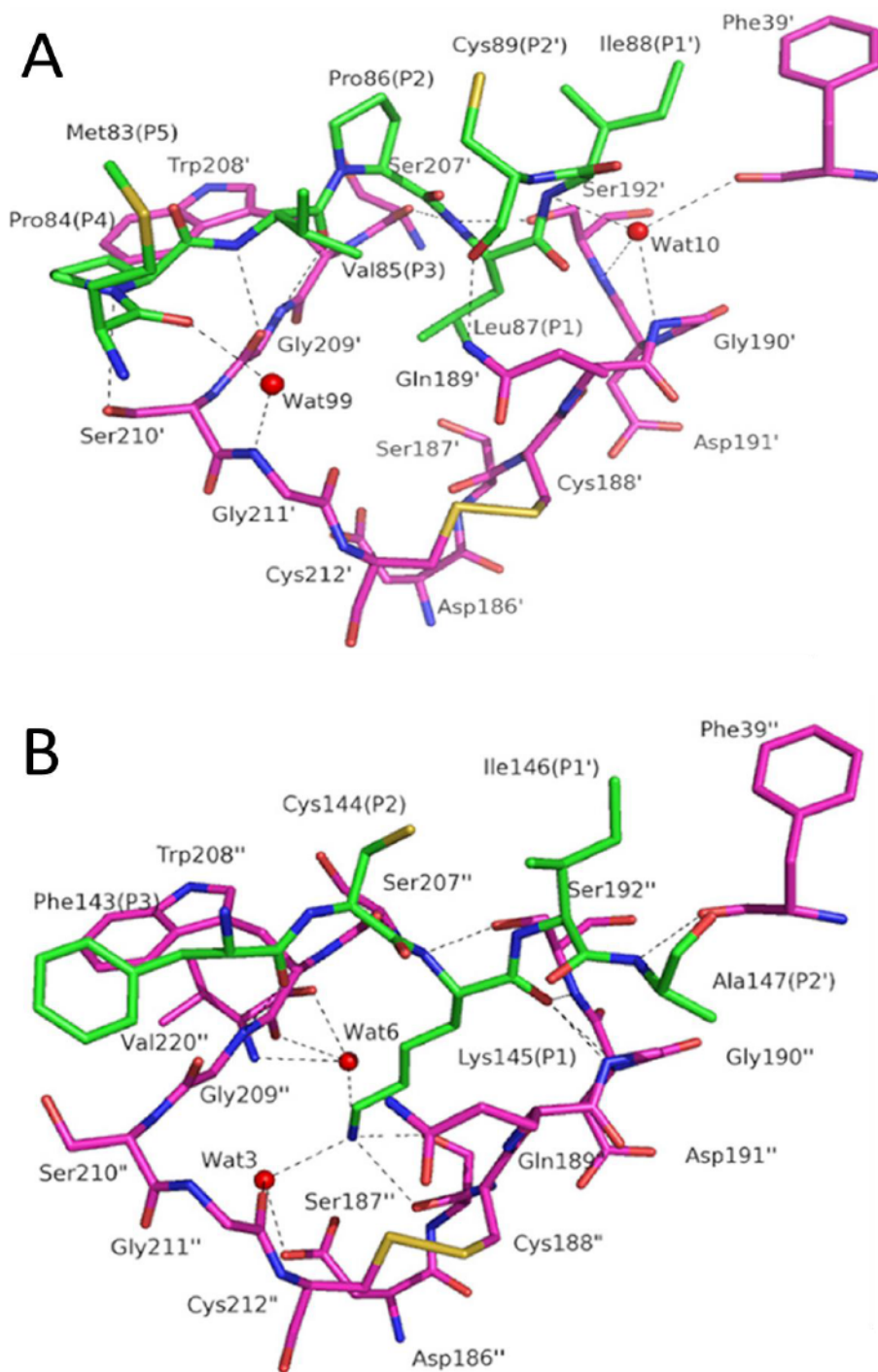


Figure 8. The views of the two interfaces. **A** and **B**, represent interfaces 1 and 2 between reactive sites of API-A (in green) and trypsin (in magenta) (35). The residues of trypsin-1 and -2 are labeled with a prime and a double prime, respectively. All hydrogen bonds are shown as dashed lines. The water molecules (Wat³, Wat⁶, and Wat⁹⁹) are shown as red spheres.

For the interface 1, the P1 residue Leu⁸⁷ snugly fits into the S1 pocket of trypsin-1. The formation of a typical intermolecular antiparallel β -sheet, which is stabilized by three main chain hydrogen bonds, also helps the binding conformation at interface 1. Moreover, the residues surrounding RS1 also contribute to the enzyme-inhibitor interaction at interface 1, via hydrophobic and hydrophilic interactions. For the interface 2, there are classic protease-inhibitor interactions presented at the interface between RS2 and trypsin-2. In accordance with the canonical binding mode, the carbonyl group of the scissile peptide bond between P1 Lys¹⁴⁵ and P1' Ile¹⁴⁶, is snugly embedded in the oxyanion hole. The main chain hydrogen bonds are mainly contributed by P1 Lys¹⁴⁵ (Figure 8) (36). In addition to direct hydrogen bonds, the N ζ atom of P1 Lys145 forms five indirect hydrogen bonds mediated by two water molecules Wat³ and Wat⁶.

1.5 Problematics

We have mentioned the essential role of proteases in cell physiology, and its relationship with several pathological conditions such as cancer, neurodegenerative disorders, and inflammatory and cardiovascular diseases. Many proteases are a major focus of attention for the pharmaceutical industry as potential drug targets against these diseases, or as diagnostic and prognostic biomarkers (1, 5).

In the clinical applications against diseases caused by proteases, protease inhibitors are often utilized as treatments to bind the target protease in a substrate-like manner, which can inactivate or control the proteases. Many protease inhibitors have been proved to have the potential of treating human diseases (22-25). It is well known that many of the protease inhibitors contain multiple homologous inhibitor domains in a single polypeptide chain. These domains are not identical to each other, but are functional inhibitors in their own right (20). Accordingly, the identification of the possible inhibitor units (domains) in a certain inhibitor can help finding out its potential target proteases. This would have a significant impact in the clinical application of this inhibitor.

As one of the Kunitz-type serine protease inhibitors, the sequence of API-A was determined in 1992 (32). Until the year of 2009, the structural data of API-A was obtained by the method of X-ray crystallography (35), which is the only three-dimensional structure of API-A being published. However, in this report, the API-A was crystallized with trypsins, the structure obtained is the ternary structure of the API-A complex with two molecules of trypsin. It is known that the conformation of many proteins changes upon binding to a ligand (37), so we may think that because of binding to trypsins, the conformation of API-A was changed by the interactions at the active sites between API-A and trypsins. Owing to the conformational changes, it could be difficult to recognize the other possible inhibitor units of API-A. Because of this, it will be very interesting to know the structural conformation of API-A in its apoenzyme form. Maybe some other possible active sites will be found, which could help in finding other potential target proteases or some other possible substrate binding patterns.

Secondly, after knowing the structural conformation of API-A in its apoenzyme form, it will be interesting to compare the conformations both in its apo-protein form and in complex. Knowing the conformational changes of API-A when it binds to the proteases can be helpful to the studies of protein-protein interactions, and also be useful in designing and developing other therapeutic inhibitors of the same catalytic mechanism.

1.6 Research objective

To better understand the principle of protein crystallization, as well as to improve the crystallization success, different protein crystallization methods may be used, such as the methods of relative crystallizability and composition modification.

In order to obtain the conformational information of the apo form of API-A, we have chosen X-ray crystallography to determine its structure. To produce a crystal of apo API-A with adequate quality for X-ray crystallography, various methods of protein crystallogenesis were used. Once we obtain the qualified crystal, through X-ray diffraction, structural data can be

collected. Finally, the data processing steps such as the phase determination, model building and refinement, are used to obtain the three-dimensional structure of apo form of API-A.

Chapter 2

Materials and methods

2.1 Methods in protein preparation

The first step in a project of protein structure determination by X-ray diffraction is to obtain a highly pure and homogeneous sample of a solution of that protein. As relatively large quantities, normally a few milligrams of proteins, are needed to form crystals, and a proper concentration of the protein solution is important to produce crystals, that protein is usually overexpressed before its purification.

2.1.1 Cell culture

The cells expressing API-A in *Escherichia coli* BL21 (DE3) were provided by the laboratory of professor Cheng-Wu Chi, Shanghai Institutes for Biological Sciences (Figure 9). The target gene of API-A was cloned into a pET28a-derived expression vector encoding 6-histidine residues (His₆ tag at the N-terminus) after the start codon. The constructs were co-transformed with PKY206, a plasmid containing the *groESL* genes of *E. coli*, encoding the chaperones GroEL and GroES (35), in order to optimize the correct folding of API-A.

The cultured cells were grown at 310 K, in 2×YT medium (5 g of NaCl, 16 g of bactotryptone, 10 g of yeast extract in 1 liter of H₂O), containing kanamycin (SIGMA) and tetracycline (SIGMA) that added at a final concentration of 30 µg/ml and 10 µg/ml respectively. When the culture had reached an OD_{600 nm} of 0.7, protein expression was induced by adding isopropyl 1-thio-β-D-galactopyranoside to a final concentration of 0.2 mM, and the culture was then incubated at 16 °C for 20 hours (35).

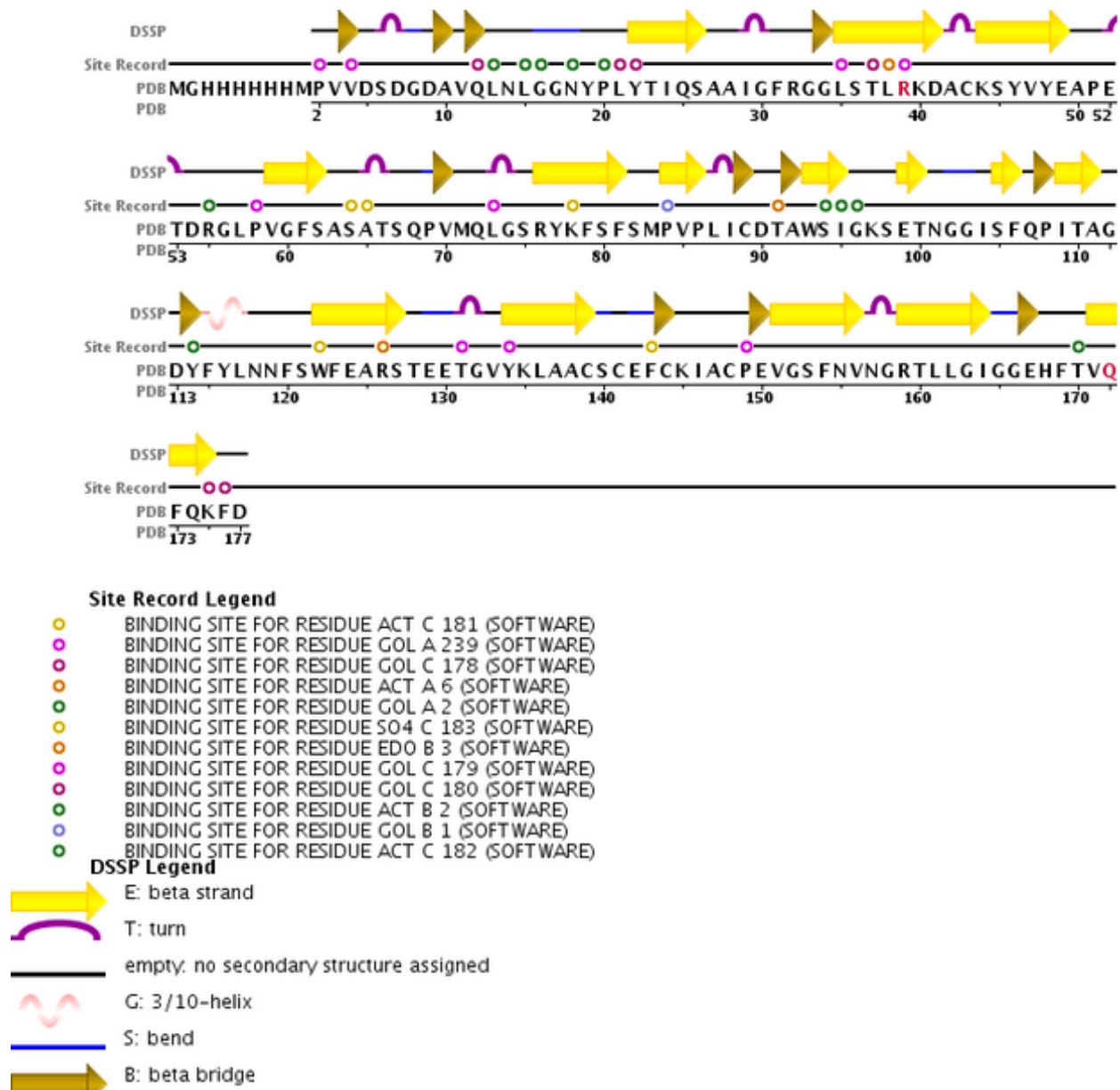


Figure 9. Sequence display of API-A, from Protein Data Bank (35).

2.1.2 Protein purification

All the procedures were carried out at 4 °C or on ice unless otherwise specified. Cells (8g) were collected from cell culture, centrifuged at 3,400 rpm, 4 °C for 20 min, (Thermo Scientific HLR-6 rotor, Sorvall RC-3B Plus Refrigerated Centrifuge) and re-suspended in the buffer A containing 300 mM NaCl, 50 mM Tris-HCl, at pH 7.5, 10% glycerol, 0.5 mM PMSF, 10 mM β-Me. Cell

lysis was carried out as follows: after adding 1 mg/ml lysozyme (from chicken egg white, Sigma-Aldrich), 20 mM imidazole and 1 µg/ml of each of the following protease inhibitors (leapetin, chymostain, antipan, aprotinin and pepstatine A), the mixture was left on ice for 15 min before sonication. The sonication was carried on ice with ten 8-s bursts separated by 10-s intervals at output 2.5, using a Sonic Dismembrator (Model F550, Fisher Scientific).

The homogenates after sonication were centrifuged for 45 min at 42,000 rpm (70 Ti rotor, Beckman), at 4 °C. The supernatants were collected and were filtered through syringe filter (0.2 µm pore size, Filtropur S), and then mixed with 5 ml of the nickel-nitrilotriacetic acid affinity resin (Qiagen) pre-equilibrated with buffer B (buffer A containing 20 mM imidazole), and incubated by rotating for 4 h at 4 °C. Then the mixture was loaded onto a column and the flowthrough was collected. The column was washed with 10 column volumes of buffer B, and 10 column volumes of buffer C (buffer A containing 30 mM imidazole). Bound proteins were eluted with buffer D (buffer A containing 300 mM imidazole).

The fractions of high protein concentration were collected into a total volume of 1 ml, using an Amicon Ultra 10-kDa cut-off concentrator (Millipore). This small volume is suitable for the next purification step by gel filtration chromatography. The Hiload 16/60 Superdex 75 (Amersham Biosciences) column was chosen. Before use, the column was equilibrated with buffer E (containing 50 mM Tris-HCl, at pH 7.5, 10% glycerol, 0.5 mM PMSF, 10 mM β-Me). A flow rate of 1 ml/min was used for both sample loading and washing.

Blue sepharose was chosen as the last purification step. After equilibration in buffer E, the sample protein was loaded at the flow rate of 1 ml/min and the column was washed with the same buffer until the optical density baseline was stable. After the unbound proteins were washed out, the bound sample was eluted using an increasing linear salt gradient, from 100% buffer E to 100% eluent buffer F (buffer E containing 500 mM NaCl).

2.1.3 Measurement of protein concentration

The Bradford assay was chosen to determine the protein concentration. The Bradford assay is a colorimetric protein assay based on an absorbance shift of the dye Coomassie Brilliant Blue G-250. By binding the sample protein, the dye of red form is converted into its blue form. During the formation of this complex, two types of bond interactions take place: the red form of Coomassie dye first donates its free electron to the ionizable groups on the protein, which disrupt the protein's native state, consequently exposing its hydrophobic pockets. These pockets in the protein's tertiary structure bind non-covalently to the non-polar region of the dye via van der Waals forces, positioning the positive amine groups close to the negative charge of the dye. The bond is further strengthened by the ionic interaction between the two. The binding of the protein stabilizes the blue form of the Coomassie dye. Thus the amount of the complex present in solution is a measure for the protein concentration, and can be estimated by use of an absorbance reader. The bound form of the dye has an absorption spectrum maximum at 595 nm. The increase of absorbance at 595 nm is proportional to the amount of bound dye, and thus to the amount (concentration) of protein present in the sample.

The Bradford protein assay is less susceptible to interference by various chemicals that may be present in protein samples. However, detergents such as sodium dodecyl sulfate and triton x-100 can interfere with the assay, as well as strongly alkaline solutions (38).

In this study, all the procedures were almost free of the use of detergents. A Beckman DU-70 spectrophotometer was used for the measurements of protein concentration. BSA was used as the standard.

2.1.4 Mass spectrometer

As a method for protein identification, we used the tandem mass spectrometry (MS/MS) identification, which is nowadays a well established method (39) (Figure 10). Being widely used to measure the characteristics of molecules, it converts them to ions so that they can be moved

about and manipulated by external electric and magnetic fields.

The mass spectrometer is composed of three elements: firstly, an ion source, such as a high energy electron beam, which causes the sample protein ionized to cations by loss of an electron; two techniques are commonly used in this step, the electrospray ionization (ESI) or matrix-assisted laser desorption/ionization (MALDI). Then through the mass analyzer, the ions are sorted and separated according to their mass and charge. This step is achieved by accelerating and focusing the ions in a beam, which is then bent by an external magnetic field. Two mass analyses are used, either “in space” or “in time.” For “in space” configurations, such as triple quadrupole (TQ), quadrupole/time-of-flight (Q-TOF), or time-of-flight/time-of-flight (TOF–TOF), the primary and secondary analyses are performed sequentially as ions travel through the instrument. For “in time” configurations, such as quadrupole ion trap (Q-IT), they are performed consecutively within the same analyzer. Finally the separated ions are measured by a detector, and a chart displaying the results is obtained (40).

A MS/MS spectrum is usually presented as a vertical bar graph, each bar represents an ion which has a specific mass-to-charge ratio (m/z), and the length of the bar indicates the relative abundance of the ion. Most of the ions formed in a mass spectrometer have a single charge, so the m/z value is equivalent to mass itself.

The spectrum can then be interpreted and correlated with theoretical peptide sequences from protein or genomic databases. The last step is to combine the peptide identification results into a list of proteins that are most likely present in the sample.

For the preparation of the sample protein API-A, the first step in this study was to reduce the complexity of the sample protein by applying protein separation techniques. The system of Fast-Performance Liquid Chromatography (FPLC) was combined with three steps of purification: by affinity with a nickel-nitrilotriacetate resin; by size-exclusion chromatography and blue sepharose chromatography. The purified proteins were sent to the “Plate-forme protéomique du Centre de génomique de Québec”, where they were cleaved into peptides using trypsin as the proteolytic enzyme, which cleaves peptides at the C-terminal side of arginine and lysine residues.

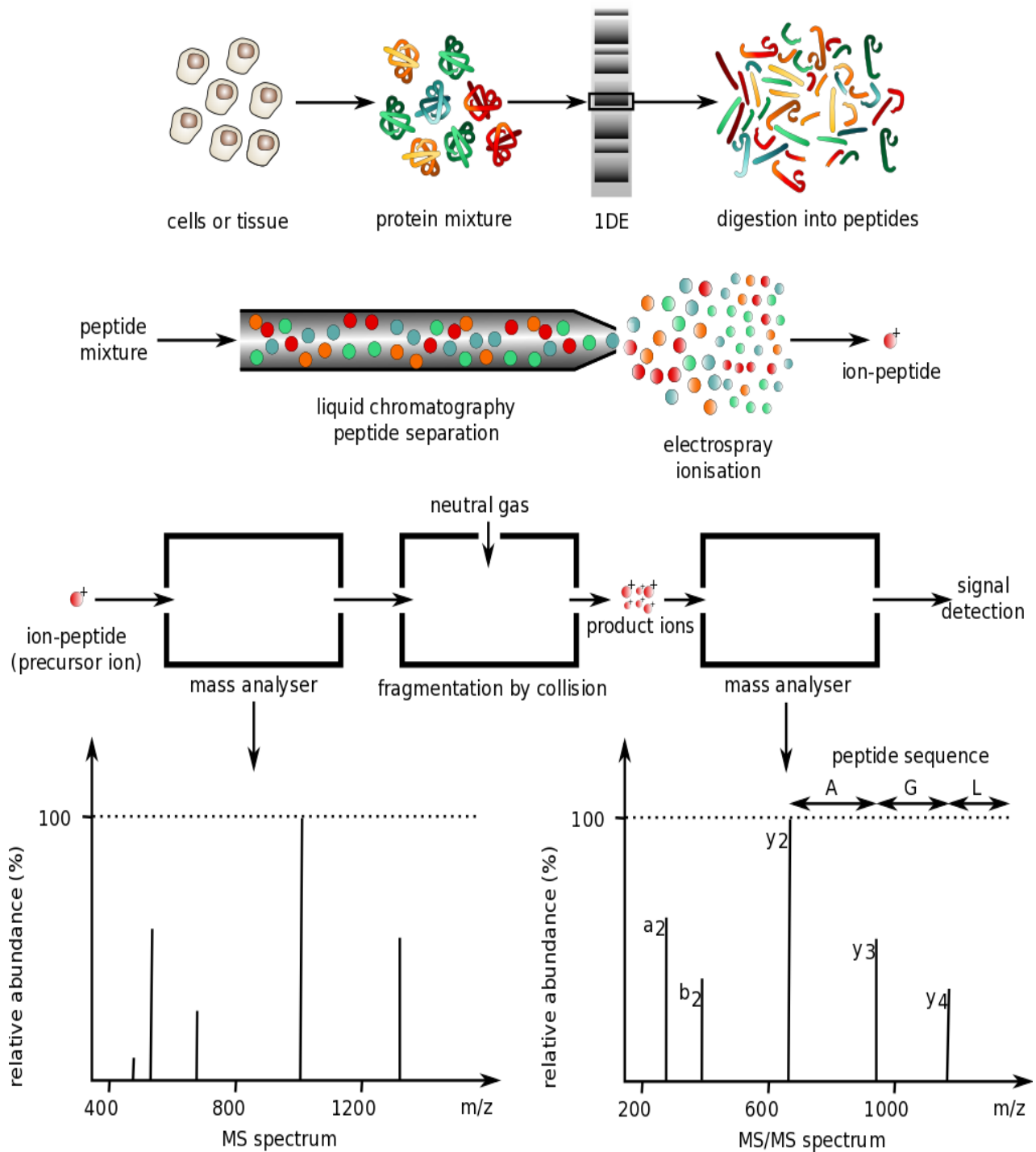


Figure 10. Mass spectrometry protocol. (Emmanuel Barillot, Laurence Calzone, Philippe Hupé, Jean-Philippe Vert, Andrei Zinovyev, Computational Systems Biology of Cancer Chapman & Hall/CRC Mathematical & Computational Biology , 2012)

2.2 Methods used for protein crystallization

In this and the following (2.3) sections, I will review all methodology available today for protein crystallization and then provide the details on methodology used in my study in the Results section (Chapter 3).

In order to produce good diffraction-quality crystals, the techniques are different for small molecules and macromolecules, owing to the differences in conformational freedom degree. For small molecules, as they tend to have few degrees of conformational freedom, they are normally crystallized by a few methods such as vapor deposition and recrystallization. On the contrary, macromolecules tend to have more degrees of conformational freedom; accordingly, the crystallization methods for macromolecules must also maintain their structural stability, for example by using crystallization conditions that will keep the tertiary structure of the macromolecules intact.

For protein crystallization, the procedures are often carried out in solutions; by lowering the solubility of target macromolecules very gradually, the production of crystallites can be favored. However precipitation is likely to happen when the process is too quick, yielding useless dust or amorphous gel. The purpose of any crystallization method is to drive the macromolecule solution from its undersaturation status to a high supersaturation status; as a result, crystal nucleation happens (Figure 11). The crystal growth can be characterized into two steps, the nucleation of crystallites, followed by the growth of the crystallites. Normally the solution condition which favors the nucleation step does not favor the subsequent growth. In the aspect of X-ray crystallography, as our goal is to obtain a single crystal, well ordered and large enough, starting with a limited amount of protein molecules, the ideal condition will be the one that favors the crystallite growth but does not favor the nucleation. If we can produce fewer crystal nuclei, we will have better chances to obtain a larger size crystal. By using an adequate crystallization method under the suitable condition, the crystallite will form first, then, along with the crystals growth, both the concentration of soluble macromolecules and the supersaturation will start to

decrease. Finally until the equilibrium has been reached, the concentration of soluble macromolecules will decrease and reach the solubility curve, thus preventing further growth of the crystals. If the method and condition are well suitable, diffraction-quality crystals can be obtained.

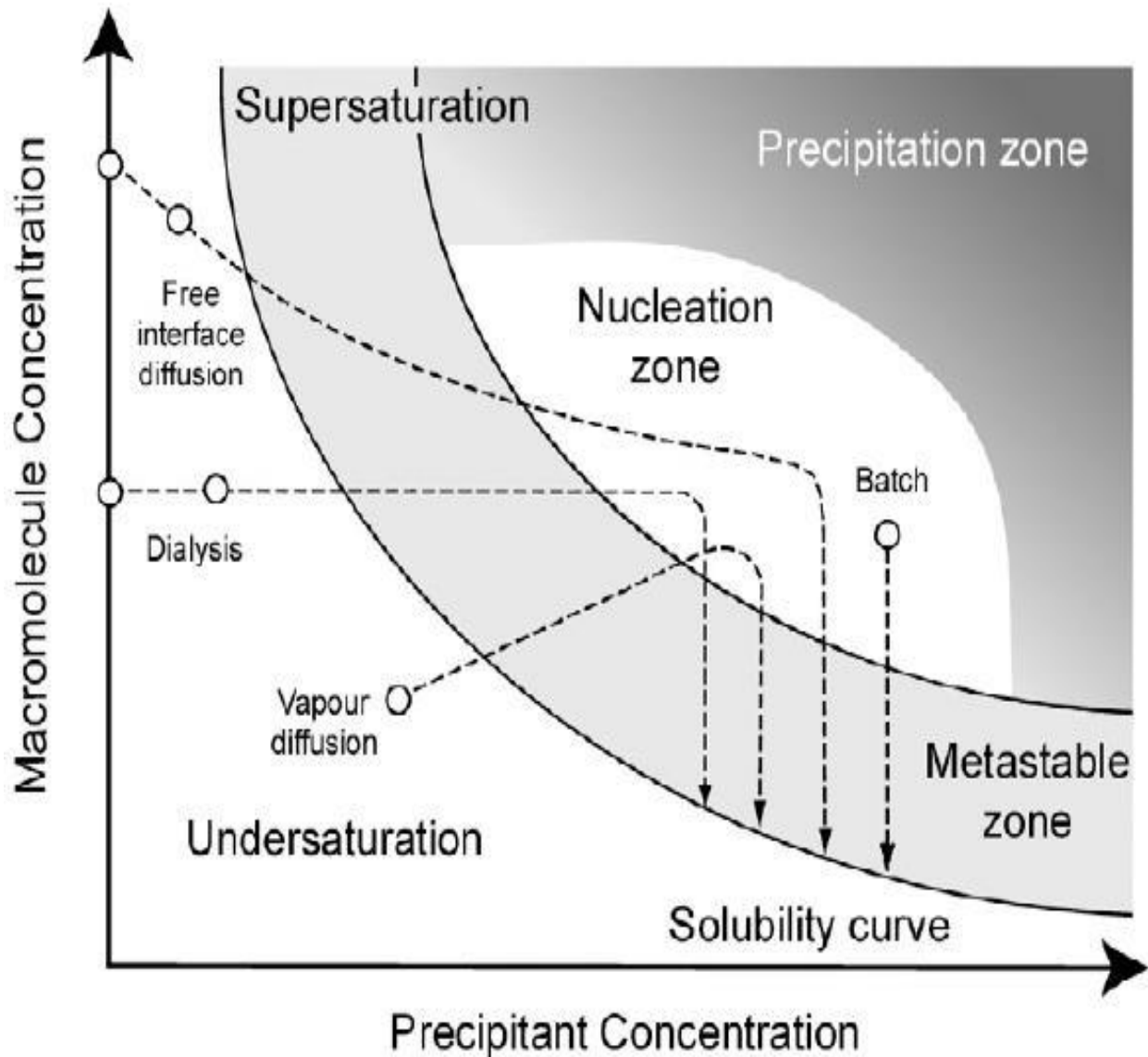


Figure 11. Equilibration pathways in various protein crystallization methods. This theoretical two-dimensional phase diagram displays how supersaturation is reached to trigger crystallization. These are based on the batch, dialysis, free-interface diffusion or vapor diffusion principles (41, 42).

The crystallization of biological macromolecules can be considered as a two-stage process. The first stage is called “screening”, which identifies certain chemical and physical conditions, under which the sample protein has a propensity to crystallize. This can be judged by the appearance of small crystallites. The second stage is called “optimization”, which is about to refine the chemical and physical parameters by different crystallization methods to produce well qualified crystals suitable for analysis by X-ray diffraction.

2.2.1 Preparation of protein solutions before crystallization

Before beginning crystallization, the sample protein needs to be concentrated and transferred to dilute buffer containing little or no salt depending on the nature of the protein. This is achieved using centrifugal concentrators. In order to screen a reasonable number of conditions, at least 200 µl of protein at adequate concentration is needed. Most proteins can be crystallized at the concentration of 10 mg/ml, so normally a sample protein that is ready to be crystallized will be concentrated to 10 mg/ml; however the sample protein concentration can also be adjusted according to the protein nature. The sample protein solution needs to be centrifuged before use, to remove the precipitated protein and leave the soluble protein molecules.

2.2.2 Vapor diffusion methods

Among the crystallization methods that are still widely used today, probably the vapour diffusion techniques are the most popular throughout the world (43, 44).

Following the same crystallization principle, three simple and practical vapour diffusion methods are nowadays widely applied (Figure 12). They were named the hanging drop, sitting drop and sandwich drop vapour diffusion method.

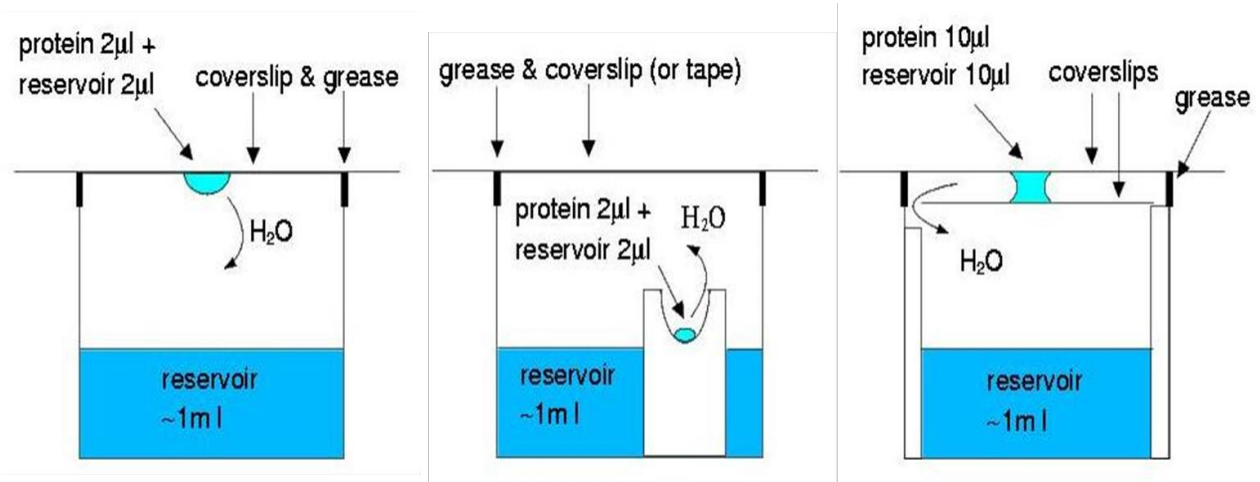


Figure 12. Schematic representation of hanging drop, sitting drop and sandwich drop vapor diffusion method. (Cited from: Methods of Protein Crystallization, 1999-2005 Airlie J McCoy, University of Cambridge)

When using vapor diffusion methods, they are usually carried out in a 24-well tissue culture plate sealed with siliconized glass cover slips. In each well, there is normally 1 ml of reservoir solution, which contains buffer, salt, crystallizing agent and additives. Very small amount of protein solution and reservoir solution are mixed by gently drawing and expelling by a pipette to form a crystallization droplet; for hanging drop diffusion method, the ideal droplet is smaller than 5 μl (45). Then, each well is closed by sealing the glass slips onto the wells using petroleum jelly. The plate is left undisturbed at operating temperature so that vapor diffusion may start to take place. The culture plate can be checked after certain periods to see how the crystals grow. The trays are placed against a dark background, and illuminated from the side, using a microscope for visual examination.

Both the crystallization droplet containing the macromolecule and the reservoir solution are at the concentration lower than that are required for the formation of crystals. In such close environment, this droplet is equilibrated against the reservoir solution, in which the crystallizing agent is at a higher concentration than it is in the droplet. Under such a condition, the equilibration proceeds slowly until the vapor pressure in the droplet equals the pressure of the reservoir solution. Along with the process, the volume of droplet decreases, as a result, the

concentration of all constituents in this drop increase to reach supersaturation (43). Moreover, during the process of equilibration, the evaporation rates from hanging drops have been experimentally determined by using crystallizing or dehydrating agents, such as ammonium sulfate, sodium chloride, PEG or MPD (46, 47). The main parameters determining the water equilibration rate are initial drop volume, water pressure of the reservoir, temperature and the chemical nature of the crystallizing agent (43).

In some studies, it is reported that crystal growth may be favoured in the slow process that permits attain the supersaturation without causing precipitation. This could be a good reason for the crystallization successes with the widely used PEG owing to its particularly slow equilibration rates (48).

2.2.3 Other methods for crystallization

There are several physical and chemical parameters that can greatly influence the results of crystallization, such as temperature, electric or magnetic field, pH, type of buffer, additives and precipitants. Owing to the fact that parameters influence the appearance and growth of protein crystals, and that every protein has its own distinctive chemical properties, the process of protein crystallization cannot be linear. After getting information from many of the steps, constant re-evaluations are required for this procedure in an iterative manner.

Methods that control or alter parameters as a function of time, pH or temperature are considered to be effective options (45, 49), in addition to the basic method of sparse matrix screening for crystallization of proteins. The latter is an approach using a set of solutions (cocktails) to screen and identify crystallization conditions. These solutions are readily available in commercial kits. These cocktails cover a range of pH, chemical species, and concentrations by varying buffers, salts, precipitant and detergents in the reservoir solution (50, 51). A sparse matrix design is used to formulate a reasonable number of cocktails (reservoir solutions) to survey the vast chemical landscape.

Protein solubility screens may be carried out first, in order to select the appropriate protein concentration for crystallization. They can probe the solubility of the proteins at different concentrations, in different buffers conditions, with different additives or stabilizers, and with different precipitating agents. Usually, proteins do not crystallize because of the lack of conformational homogeneity of the purified protein. Protein solubility screens can function as the optimum solubility screen, to obtain the most homogeneous and monodisperse protein conditions; it is helpful for proteins that usually aggregate and cannot be concentrated before setting up crystallization screens. In most cases, crystals appear in the precipitate cloud, because crystals can only grow from supersaturated solutions; in other words, the solution that contains precipitates can probably bear crystals.

After protein solubility screen, fast screening of crystallization conditions can be carried out. These commercial available sparse matrix or grid screen solution kits can be obtained from many companies, such as sparse matrix, matrix screen PEG/ions, grid screen ammonium sulfate, grid screen PEGs, grid screen PEG/LiCl, grid screen alcohols, grid screen salts.

A grid or a matrix screen for protein solubility can also be designed using commercial kits. By observing the precipitant situation, the modification trend can be judged considering that protein solubility decreases proportionally as the polymer concentration increases. In the case of clear drops, the concentration of precipitant agent can be increased to produce a stronger precipitating environment. When using PEG as precipitant agent, changing to larger molecular weight can be advisable. Reversely, in the case of heavy precipitate, lower concentration and smaller molecular weight precipitant agent can be adopted. Salt and pH can be adjusted to reach a condition which produces the environment suitable to make the switch between clear drop and heavy precipitate. Normally in this step, at least some microcrystals, polycrystalline aggregates or thin plates can be observed (Figure 13).

Detergents are often used as additives in crystallization trials. In the protein solutions, the hydrophobic interactions may lead to nonspecific aggregation and consequently restrain crystallization. Some small amphiphiles can be used as additives, such as benzamidine, ethanol,

dioxane, 1,6-hexanediol, ethylene glycol and butyl ether, for their ability to influence micelle stability (52). Considering that the length of the detergent chain may influence the micelle size, the amount of precipitant can be decreased by increasing the length of the detergent chain. The advantages of adding additives is that they can form hydrogen bonds or electrostatic, reversible crosslinks between proteins in the crystal lattice, to cause consequent favorable changes in its physico-chemical properties or conformation (45). In the experiment shown in Figure 13, 72 additive conditions of Additive Screen (Hampton Research) were applied in the crystallization drop to improve the crystal growth and quality.

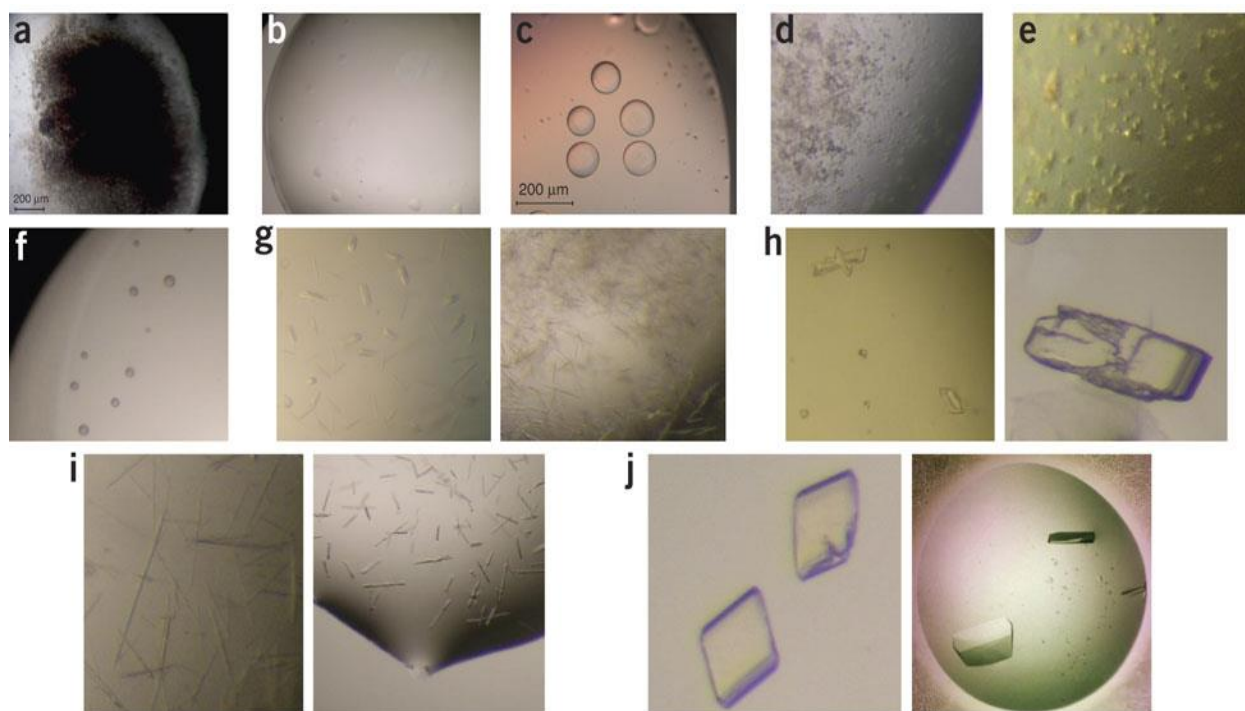


Figure 13. Classification scheme for the results of a crystallization screen (53, 54). (a) Heavy brown or flocculent precipitate, (b) clear drop, (c) phase separation or light precipitate, (d) granular precipitate, (e) microcrystalline precipitate, (f) spherulites, (g) micro-crystals, (h) multiple crystals, (i) small crystals (needles, thin plates), (j) well shaped single crystals.

Other methods can be used supplementarily in different cases, such as crystallization in electric or magnetic fields (53, 54), under pressure (55), under micro- or super gravity (41).

2.2.4 New ideas in crystallization

2.2.4.1 Relative crystallizability

“Relative Crystallizability” was used as a new measurable parameter to semi-quantitatively evaluate the possibility of protein crystal growth (PCG) under certain conditions, such as temperature, pH and salts (56, 57). It is defined as the percentage of the nucleation zone over the phase area delineated by the experimental protein and precipitating agent concentration ranges in the two-dimensional-phase diagram, since spontaneous nucleation occurs in the area of nucleation zone. Crystal-solution phase diagrams are designed to plot the initial concentration of protein versus the concentration of precipitating agent, under certain crystallization condition. The solutions were placed in incubators at different temperatures (288, 295, 300, and 303K).

By confirming the crystal growth situation in Sparse Matrix Screen experiments, which use PEGs as precipitating agents, the proteins crystal growth is in excellent agreement with the relative crystallizability. The relationship between solubility dependence, relative crystallizability and crystallization success has been demonstrated, which could be used to identify efficient crystallization regions and to provide an efficient approach to PCG (Figure 14).

In this study, by examining a target protein and drawing out the nucleation zone area (to which the modification can be realized by modifying temperatures or other parameters in order to affect the zone), the authors (Zhu et al, 2006) discovered a rational approach to crystal growth as well as for selecting the crystal form. For example, the modification as a function of temperature can give a significant change to the phase diagrams, shedding light to find a suitable temperature with a higher probability of obtaining new crystals.

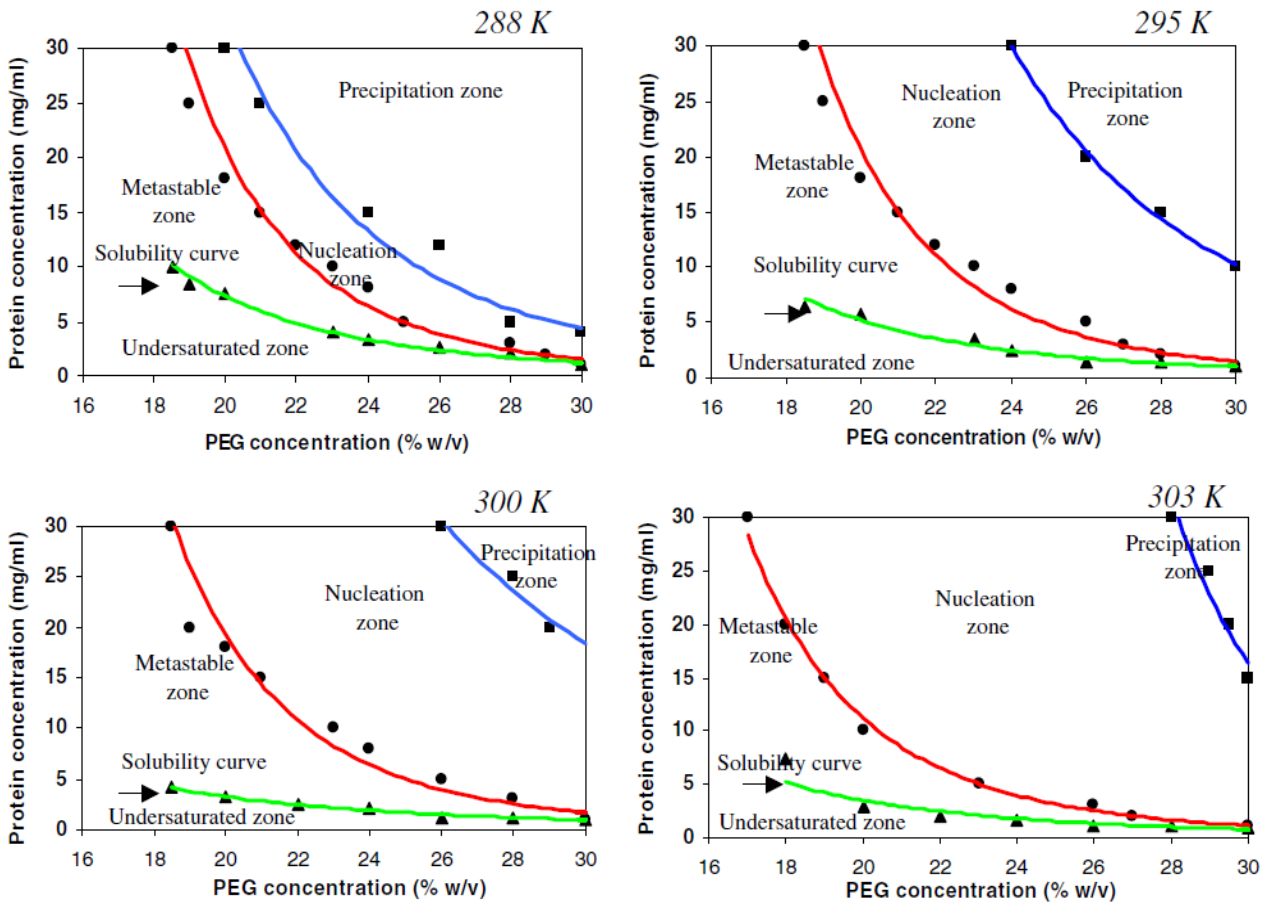


Figure 14. Two-dimensional-phase diagrams of protein concentration versus PEG concentration (56). Solubility curve (in green) was determined from the residual concentration in equilibrium with crystals 50-days after the initiation of crystallization at various temperatures. Nucleation and precipitation data are plotted in red and blue, respectively. In this diagram the precipitation zone is denoted by the region above the precipitation curve while the nucleation zone, where spontaneous nucleation occurs, lies between the nucleation and precipitation curves. The metastable zone is situated between the nucleation and solubility curves where crystals grow, while the area below the solubility curve is the under-saturated zone.

2.2.4.2 Composition modification

For the success of crystallization analysis and statistics, a method focusing on the composition modification of protein and precipitant mixing ratio was found, in the hope to increase the success of crystal growth (58, 59).

The composition modification is often used as an optimization method, after an initial condition

was identified through screening; it consists in varying the concentrations of the macromolecule and of the precipitant in a systematic manner. This strategy was implemented by varying the volume ratio of protein sample to the reservoir solution in the drop to initiate crystallization, aiming to reach the nucleation zone (Figures 15 and 16); this method was reported to be an effective and efficient way to produce high-quality crystals using batch methods (60).

By the means of composition modification, the influence of several factors can be studied: for example the macromolecule concentration, precipitant concentration, salt. Accordingly, several different conditions are produced and can be tested as screening approaches. Also, a certain range of pH values can also be scanned, because the protein and reservoir solutions typically contain different buffers at different pH values; indeed, as the volume ratio of protein to reservoir solutions changes, different pH values are produced in these crystallization drops, which is effectively used as a variable during the screening experiments (59).

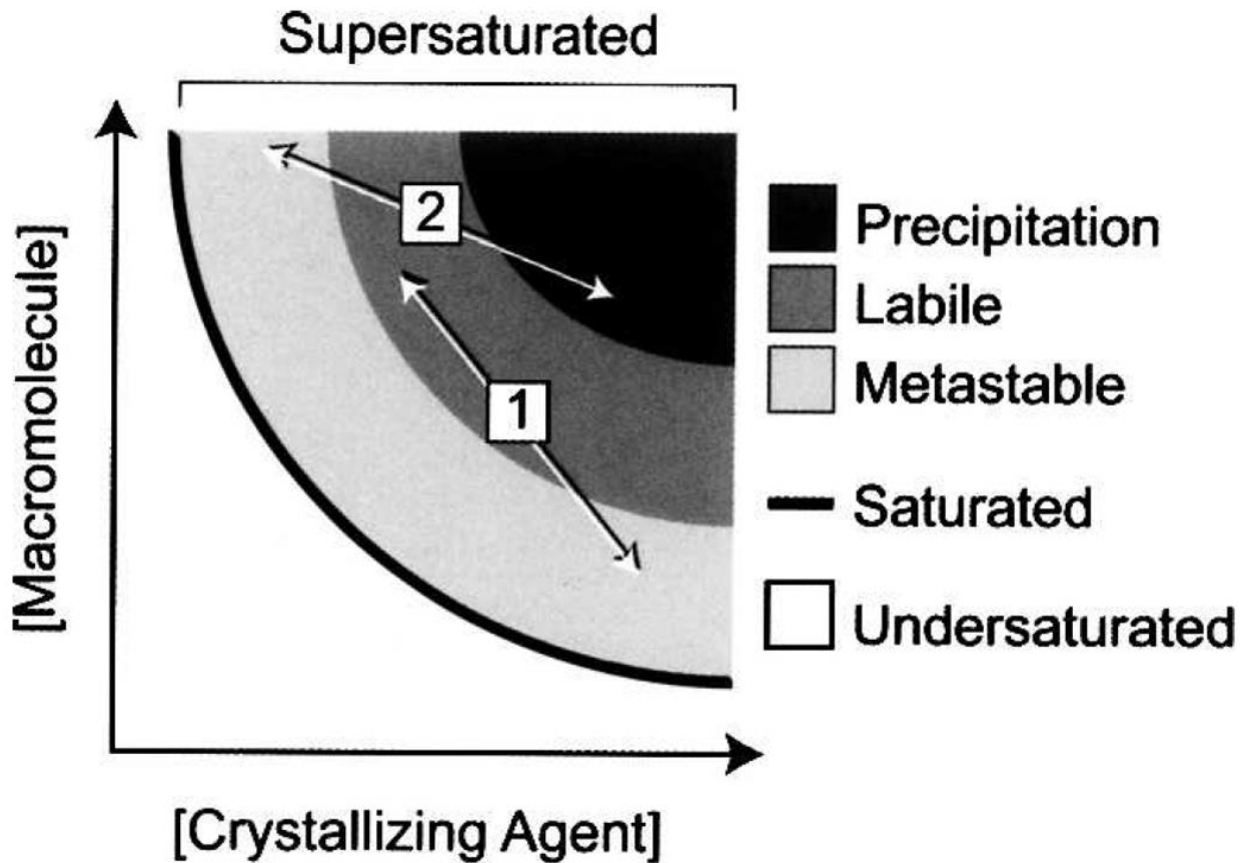
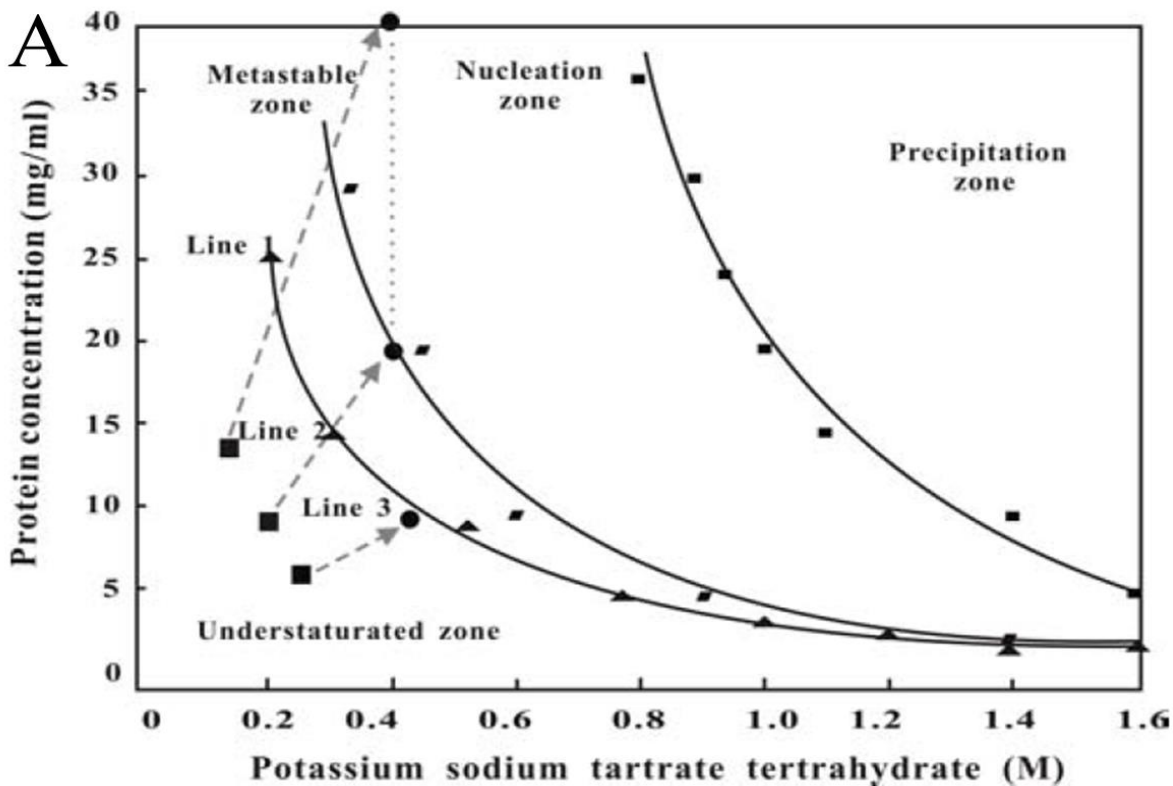


Figure 15. Drop volume ratio effect on supersaturation (59). We suppose in certain conditions of the screening trials, that crystals can be observed with a 1:1 ratio of protein to cocktail solution (reservoir solution) (1) or (2). Holding all other variables constant, we assume this experiment falls some place in the labile zone (nucleation zone) where spontaneous homogeneous nucleation will occur. Varying the volume ratio of protein to cocktail solution will sample points that lie roughly along a path indicated by the arrows on the graph. Different areas in the solubility diagram where the protein concentration is higher and the precipitating agent concentration is lower, and where the precipitating agent concentration is higher and the protein concentration is lower will be sampled.

In this study of composition modification, the crystallization “hits” are defined as the number of trials where crystals appeared, the “hits increase” represents the number of new crystallization conditions carried out by composition modification, and the “hits increase” reveals the improvement of “hits” separated from the “hits” from the initial screening (Table 2). A significant improvement in screening using this strategy was shown by statistical analysis. Carried out at different temperatures, the average improvement of “hits” is between 32 and 42%.

By using the method of composition modification, the protein crystal quality can generally be improved. Moreover, some new crystals were produced by reaching the nucleation zone which initiate crystallization. It is also demonstrated that composition modification can significantly increase crystallization success (1.3 times) after the improvement of “hits” by temperature screening (58).



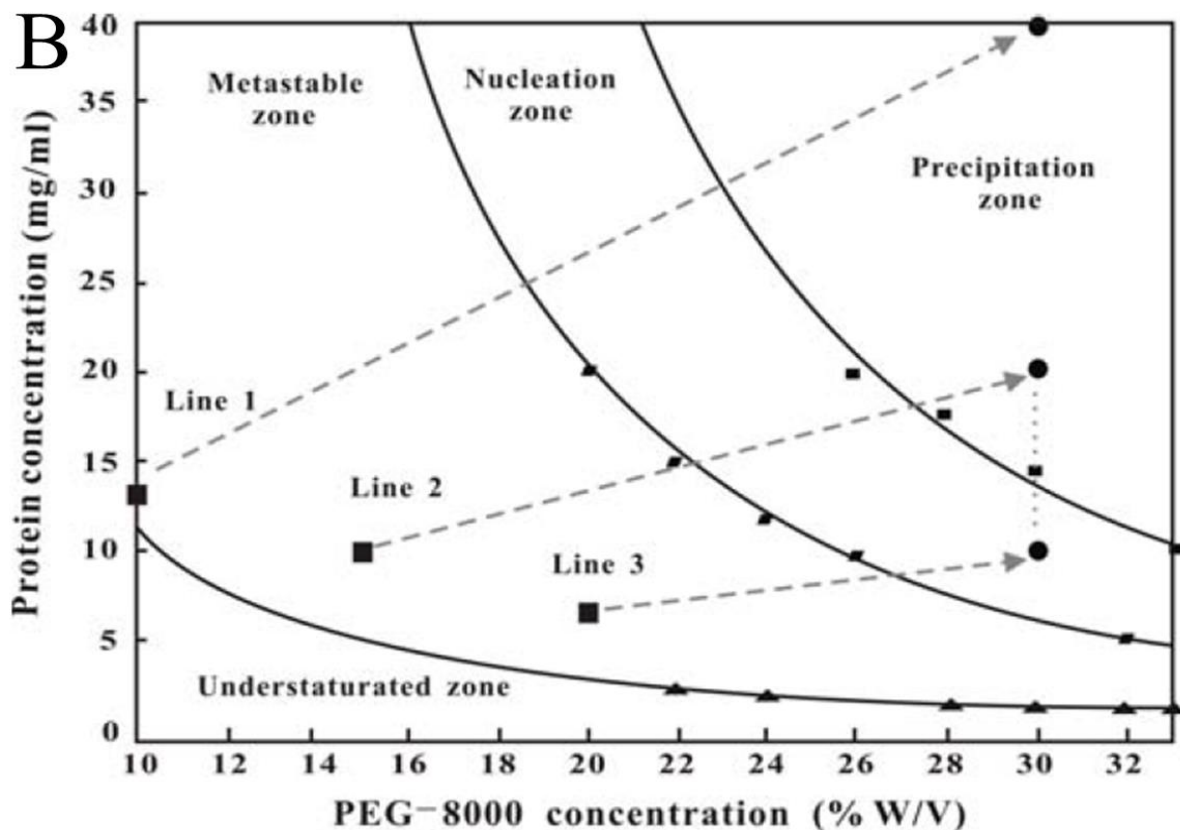


Figure 16. Two-dimensional-phase diagrams of thaumatin (A) and proteinase K (B), and prediction of trajectories of varying composition modifications at 300 K and 295 K, respectively (58). Initial screening conditions are denoted by large black squares, crystallization conditions after equilibrium by black dots. The grey broken lines show that the different trajectories terminate at the same precipitant concentration. There are three trajectories for the composition modification: the ratio of protein to precipitant under the initial condition at 1:1 (line 2), 2:1 (line 1), and 1:2 (line 3), predicted from the solution concentration equilibrium governed by precipitating agent concentration. The grey dot lines show these trajectories in this experiment. For thaumatin, according to the experimental phase diagram, it fell into the metastable zone, near the nucleation zone. The crystals were obtained after the volume ratio of protein and precipitant change to 2:1, as the drop entered into the nucleation zone. Another example is shown by proteinase K, precipitant was found in the drop with initial condition, and it lay within the precipitant zone. The crystals were obtained after modifying the volume ratio to 1:2, when the drop entered into the nucleation zone.

Table 2. “Hits” and “Hits Increase” obtained by conventional screening and composition modification (58).

proteins	277 K		288 K		295 K		300 K		summary		
	“hits” ^a	“hits increase” ^b	“hits”	“hits increase”	“hits”	“hits increase”	“hits”	“hits increase”	average “hits” improvement (% ± standard deviation) ^c	number of independent crystallization conditions by conventional screening ^e	number of independent crystallization conditions only found by composition modification ^e
lysozyme	43	8	40	14	31	14	26	13	37.2±7.7	53	13
thaumatin	3	0	3	4	1	1	2	2	83.3±57.7	6	3
catalase	37	6	30	5	31	4	30	3	13.9±3.1	47	10
trypsin	7	0	3	2	4	1	1	0	ND	7	3
ribonuclease A	2	0	6	2	5	2	5	4	38.3±32.8	13	6
ribonuclease S	11	2	6	2	7	1	9	2	22±8.2	19	4
myoglobin	1	0	2	1	1	0	3	0	12.5±25	7	1
hemoglobin	2	0	1	1	2	1	4	0	ND	9	0
NAD kinase	0	1	1	3	4	1	1	3	^d	6	2
chymotrypsinogen A	16	6	19	5	21	4	29	8	27.6±7.6	41	7
concanavalin A	33	4	34	5	33	6	38	6	15.2±2.5	52	4
proteinase K	3	4	6	9	7	3	18	9	94±55.5	25	6

a, number of crystallization conditions yielding crystals during the initial screening; b, increased number of crystallization conditions by composition modification; c, “hits” improvements were calculated by the average value of “hits increase” divided by “hits”, the values of stand deviation for some proteins are large due to their different temperature dependence; d, the data of NAD kinase was not used because there were no “hits” at control at 277 K; e, number of independent crystallization conditions at four temperatures, no repetitive crystallization conditions were included.

Twelve proteins were tested at four different temperatures using this method (Table 2). For most proteins tested, screening efficiency was improved. Lysozyme and chymotrypsinogen A have the highest “hits increase” after composition modification. New lysozyme crystals were obtained in 8, 14, 14 and 13 new conditions at 277 K, 288 K, 295 K and 300 K, respectively, after composition modification. However, some protein crystals were found by adopting the method of hits composition modification, the same proteins did not produce crystals by using conventional screening strategies. For NAD kinase, no crystal was found in conventional screen, after modifying the volume ratio of protein and reservoir solution to 1:2, crystals were found in one screening condition. Thaumatin crystals were found using 6 independent conditions by the conventional screening strategy, and new crystals were found in 3 new conditions additional using composition modification. It was the same for the ribonuclease A, 6 new conditions yielding crystals by composition modification compared to the 13 crystal conditions which adopted the conventional strategy.

The experimental data demonstrates that protein crystallizability can be increased by the approach of hitting a larger part of the nucleation zone. On the other hand, for some proteins of low solubility, this approach also facilitates the initiation of crystals, by starting from a larger volume of protein solution in the initial mixture to keep a certain amount of protein from dissolving out, which enables the weakly soluble protein gradually to become more concentrated during the vapor diffusion process and to produce crystals.

2.2.5 Preparation of protein crystals before structure determination

When a good size crystal is formed, it is important to safely transfer it for the X-ray diffraction data collection at cryogenic temperature. Indeed, cryocrystallography is essentially required nowadays to get any kind of reasonable diffraction data from protein crystals. A widely used cryocrystallography method is the flash-cool crystal mounting. By using this method, one can mount and store crystals at convenient times, in view of analyzing them days or even weeks later.

In general, the solutions that crystals are grown in are unsuitable for flash-cooling and vitrification because ice can form in these solutions during this process, owing to the presence of large solvent channels in protein crystals; solvent freezing can damage the protein crystal and decrease the resolution of the X-ray diffraction data. To avoid this problem, we replace the solution around the crystal with a cryoprotectant solution. There are two required characteristics of a cryoprotectant solution: (1) it must vitrify without forming ice and (2) it must not degrade or damage a crystal that is placed in it before the crystal is flash-cooled. When making a cryoprotectant solution, a cryoprotectant will be added to the crystal growth solution. It is important to not reduce the concentration of the precipitant and other compounds in the solution except for water. That is, to replace water with cryoprotectant and not the salt or PEG. Reasonable cryoprotectant concentrations that work well are in the 25–50% range for glycerol, ethylene glycol (v/v), for PEGs (v/v or w/v), for saturated sugars (% saturation), and alcohols (v/v) (61).

To transfer the crystal, we mount the crystal from its growth solution using fine nylon loops or glass/quartz fibers. These loops or fibers are attached to the copper mounting pin held on by a magnetic base. When a crystal is caught, it is washed quickly through the cryoprotectant solution to remove the attaching soluble protein molecules, and so the crystal will be surrounded by the cryoprotectant solution. The crystal spends very little time in the cryoprotectant before it is plunged into liquid nitrogen. Then it can be stored for weeks before data collection.

2.3 Methods in protein structure determination

2.3.1 Nuclear magnetic resonance and X-ray crystallography

To know the three-dimensional structure of a target protein is essential, because the structure has great importance both for its practical applications such as drug discovery as well as for fundamental biochemical and molecular biology studies (62). Two methods are now commonly used in determining the atomic structure for proteins: nuclear magnetic resonance (NMR) and X-ray crystallography (63-65) (Figure 17).

Nuclear magnetic resonance allows the direct, noninvasive determination of protein structure. However, it has its own disadvantages, such as low sensitivity and usually the proteins molecular weight must be less than 30,000 Da. On the other hand, X-ray crystallography is not limited in these respects, but it requires a single, highly ordered three-dimensional crystal, which imposes a preliminary work of crystallogenesis. In some cases, people may use more than one method, as NMR and X-ray crystallography are known to be complementary methods for small protein structure determination (66).

To proceed with a NMR experiment, many steps should be followed: sample preparation (including isotopic labeling), NMR data collection, resonance assignment, assembling of distance and bond angle constraints from NOE (Nuclear Overhauser Effect) and J-coupling, model building based on distance geometry calculation by restrained MD (molecular dynamics),

simulated annealing, energy minimization, finally the 3D structure will be determined (67).

For X-ray crystallography, the most precise way is to use single-crystal X-ray diffraction, where the crystal has sufficient purity and regularity. When carrying out this method, a beam of X-rays generated from a synchrotron strikes a single crystal to generate scattered beams, which will then land on detectors such as a piece of film. On the detector, a diffraction pattern of spots called reflections reveal the information of evenly spaced planes within the crystal. Along with the crystal being rotated, the intensities and angles of the diffracted beams can be recorded. These data can help us determine the mean chemical bond lengths and angles within a few thousandths of angstrom as well as within a few tenths of degree.

To carry out a single-crystal X-ray crystallographic analysis, several steps need to be followed. Firstly, the most important and difficult step, is to get a qualified crystal so that when it is hit by an X-ray beam, it can provide high resolution diffraction data which are needed to determine the three-dimensional structure of the protein (63). The crystals must be as pure as possible, and free from structural and charge heterogeneity. That is to say, pure in composition, regular in structure and possess high internal order without significant internal imperfections such as cracks or twinning. They also need to be of sufficient size (typically 0.1 mm in all dimensions). When crystallizing a protein, the purified protein molecules slowly assemble from an aqueous solution. This process is developed through adopting a consistent orientation, by aligning themselves in a repeating series of “unit cells” which is called the crystalline “lattice”, which is held together by non-covalent interactions to gradually form the crystal (65). There are several factors that require consideration to reach a well-ordered crystal, like homogeneity, concentration of protein, pH, temperature, and precipitants (45, 65). In order to reach a sufficient homogeneity, the protein usually requires at least 97% homogeneity. Moreover, pH conditions are very important, as different pH values can result in different packing orientations. Buffers, such as TRIS-HCl or HEPES, are often necessary to maintain a particular environment for the protein molecules assembling and aligning, which is closely related to packing orientations. And precipitants, such as ammonium sulfate or polyethylene glycol, are important compounds that cause the protein to

precipitate out of solution (45, 68).

Secondly, the crystal obtained is placed facing an intense monochromatic (i.e. single wavelength) X-ray beam, to obtain a regular reflection pattern. As the crystal is being gradually rotated, every orientation is recorded at which intensity of the spots differs. Multiple data sets containing reflection information may be collected.

Thirdly, with the help of complementary chemical information, and repeated computational refinement, such as data integration and scaling, phase diagram determination, model building, model refinement, finally these collected data would produce a model of the arrangement of atoms within the crystal, in other words, the 3D structure.

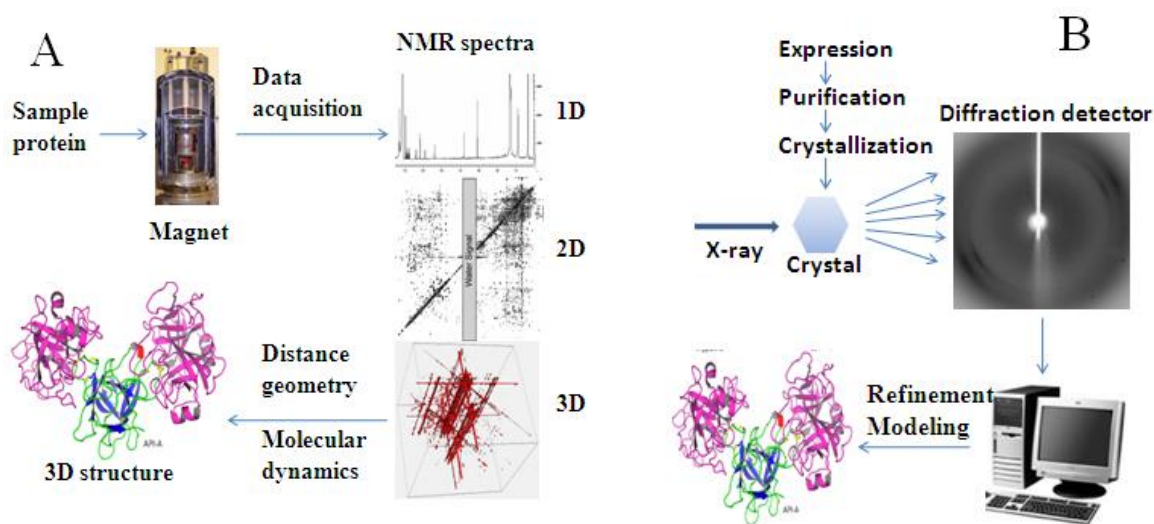


Figure 17. Process of protein structure determination by NMR (A) and X-ray crystallography (B) (45, 68).

Besides the relative merits and limits of NMR and X-ray crystallography, the total amounts of protein structures solved by these two methods respectively are different (Table 3), which is also an important factor to consider when choosing the method for target protein structure determination.

Table 3. The summary table of released entries from Protein Data Bank (PDB). Table was obtained the 7th September, 2014.

PDB Current Holdings Breakdown

Exp. Method	Proteins	Nucleic Acids	Protein/NA Complexes	Other	Total
X-RAY	85247	1556	4537	4	91344
NMR	9277	1089	216	7	10589
ELECTRON MICROSCOPY	576	67	190	0	833
HYBRID	63	3	2	1	69
other	157	4	6	13	180
Total	95320	2719	4951	25	103015

The released entries from Protein Data Bank (PDB) until the 7th September, 2014, are shown in Table 3. By data statistics, for proteins, by using X-ray crystallography, 85247 entries are released, by using NMR, 9277 entries are released. Accordingly, I assume that, by adopting X-ray crystallography to be the structure determining method for apoenzyme API-A, we may have a higher probability of success.

2.3.2 X-ray data collection

After obtaining a diffraction-quality crystal to be sent for X-ray data collection, proper care needs to be taken for safely transfer. Normally, the reservoir solution containing 15% glycerol was used as a cryoprotectant for the flash-cooled crystals in liquid nitrogen. Crystals of API-A were washed in the cryoprotectants and flash-cooled in liquid nitrogen, then they were sent to Advanced Photon Source (APS) at Argonne National Laboratory for data collection.

Diffraction data were collected using wavelength of 0.97 Å and a MAR CCD detector for data

collection and analysis. Data images were recorded while the single crystal was rotated from 0° to 180° by 1° steps.

In the X-ray diffraction pattern of a protein crystal, X-rays are absorbed as a pattern of dots, the dots normally form concentric circles of reflections. Analysis of the diffraction patterns will provide a template of the electron densities within the protein, which is essential for further structure configuration.

Chapter 3

Results and discussion

3.1 Practice in crystallization methods

3.1.1 Phase diagram study of trypsin: nucleation curves at various temperatures.

It is reported that the dependence of relative crystallizability on temperature is closely related to solubility modification. For the proteins whose solubility changes with temperature, a certain temperature can be chosen to have the largest proportion of nucleation zone over the whole phase diagram area (56). In this experiment of trypsin phase diagram (Figure 18), nucleation curves were determined at 277 K, 288 K and 295 K, and plotted by protein concentration versus PEG concentration. It is obvious that under the same crystallization condition, the nucleation curve of trypsin shifted with temperature, which demonstrates that the solubility of trypsin changes with temperatures. At the same concentration of precipitant agent (PEG), at lower temperature, less protein can be used to attain the nucleation zone; at the same concentration of protein, at the lower temperature, the nucleation zone can be reached with less precipitant agent. In other words, the nucleation curve of trypsin moves to much lower protein and precipitating agent concentrations from 295 to 277 K. For example, by using the same concentration of precipitant agent, 30% PEG (w/v), crystals can be observed in the crystallization trials with 60 mg/ml trypsin at 277 K, whereas it was at 110 mg/ml that crystals were found in the screening at 295 K. From previous reports (56, 57) and experimental data (data of precipitation curves was not given in Figure 18), the nucleation zone, which is located between the precipitation curve and the nucleation curve, increased from 295 to 277 K, it is also clear that the augmentation of the nucleation zone facilitates PCG success. In this study, it was found that within the nucleation zone, at the middle part of this zone, the biggest size, good shape and quality crystals were produced. Accordingly, to construct a phase diagram can be helpful in choosing a proper condition to obtain better crystal quality.

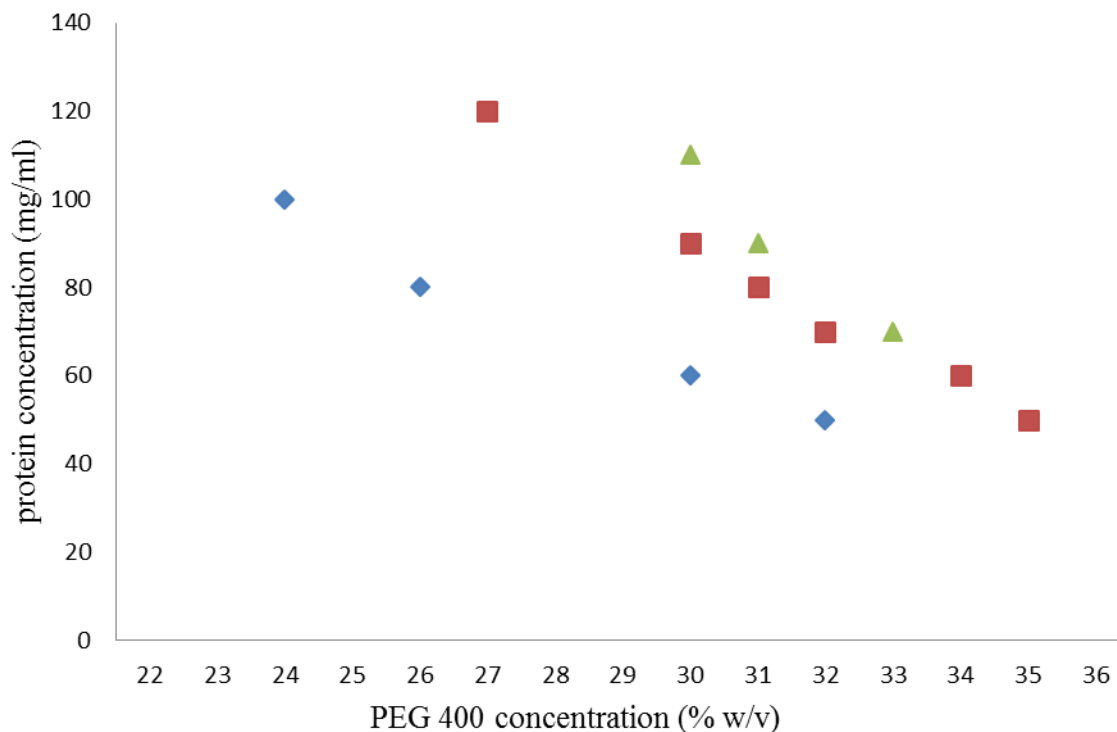


Figure 18. Phase diagram of nucleation curves of trypsin, under different crystallization temperatures. Nucleation curves were determined by protein concentration versus PEG concentration at various temperatures. Temperature of 277 K (◆), 288 K (■) and 295 K (▲) are shown, respectively. Nucleation curves were determined by crystallites in equilibrium of 50-days after the initiation of crystallization at various temperatures. In this diagram the nucleation zone is denoted by the region above the nucleation curve, where spontaneous nucleation occurs, while the metastable zone is situated below the nucleation curve where crystals continue to grow after nucleation. Crystallization condition: TRIS-HCl buffer at pH 8.5, 0.2 M CaCl₂, and PEG 400 were used in different concentrations. All crystallizations were carried out by hanging drop vapor diffusion method, which were initiated by mixing one volume of protein with one volume of reservoir solution. Crystals were observed on video images taken with an optical microscope (LEICA MZ APO) equipped with a color CCD camera (Sony).

In the later screening studies, the numbers of conditions producing protein crystals in certain screens, which are named the “hits”, used to evaluate the success rate, are in accordance with the phase diagram using temperature as a factor. In the sparse matrix screen of trypsin, for example the Nextal Classics Suite 96 conditions, at the same protein concentration and crystallization conditions, more “hits” were found at 277 K than 295 K, which is to say, for trypsin, it has decreasing “hits” dependence on temperature.

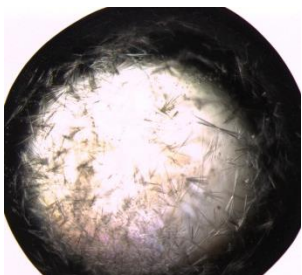
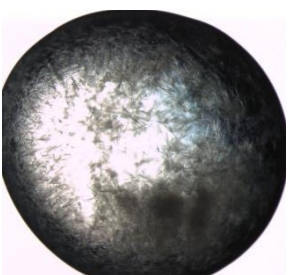
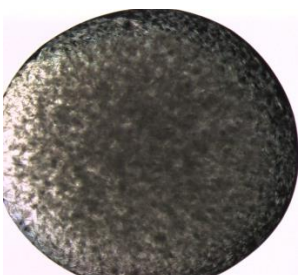
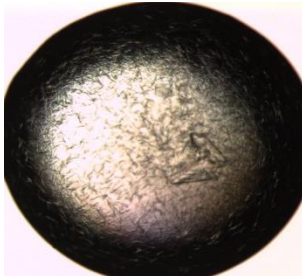
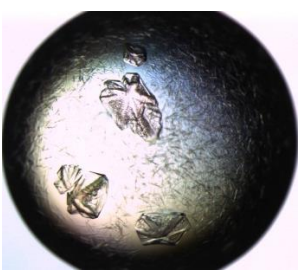
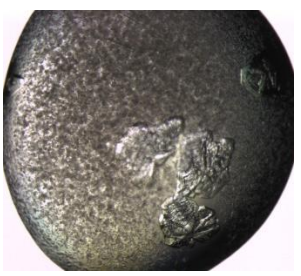
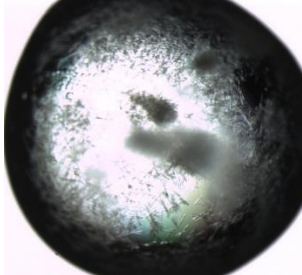
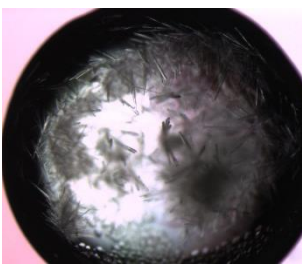
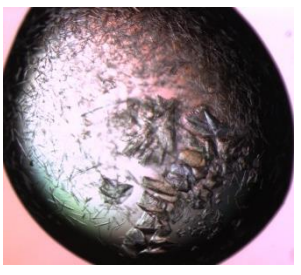
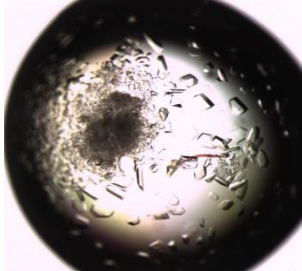
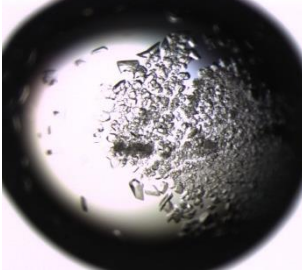
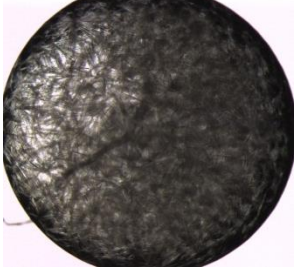
3.1.2 The effect of composition modification

The method of composition modification was implemented for two proteins (58), glucose isomerase and β -amylase, at concentrations of 30 mg/ml and of 18 mg/ml, respectively. In order to eliminate the influence of temperature in the crystallization results, all the experimental trials were carried out at a stable room temperature (Table 4, 5).

In the process of experiments, we cannot first find out the best volume ratio to produce crystals. Whereas we modify the volume ratio, for example when we produce a clear drop under a certain condition, we can adjust the drop to higher super-saturation by increasing the protein content or vice versa, to reach the optimal condition which produces crystals.

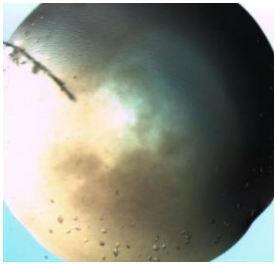
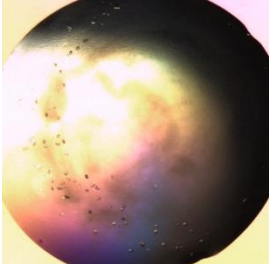
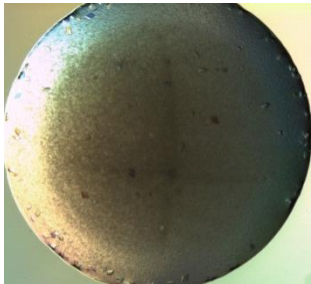
By comparing the different volume ratios in condition 1 (see legend in Table 4), the best volume ratio is 1:2, which produced the least precipitation and the best crystals, the most precipitation appeared at volume ratio 2:1 and nearly no crystal can be seen. In condition 2 (see legend in Table 4), the best size of crystals appeared at volume ratio of 1:1. In condition 3, the biggest crystals were produced at volume ratio of 2:1. In condition 4, at volume ratio of 1:2, the best size and shape of crystal can be found, at 1:1 there were too many nuclei appeared that impeded the growth of each nucleus, at 2:1 there were too many nuclei and the protein concentration was too high so that different shape of crystals were produced. In condition 5, slight precipitation appeared along with the formation of crystals, and at the volume ratio of 2:1, less nuclei were produced so that they grew bigger.

Table 4. The effect of composition modification to the crystallization of glucose isomerase.

Drop volume ratio (protein:reservoir solutions)	1:2	1:1	2:1
Condition 1			
Condition 2			
Condition 3			
Condition 4			

Crystallization conditions are chosen from Nextal Classics Suite 96 conditions. Condition 1: 0.2 M ammonium acetate, 0.1 M sodium citrate pH 5.6, 30% (w/v) PEG 4000; condition 2: 0.2 M lithium sulfate, 0.1 M TRIS-HCl pH 8.5, 30% (w/v) PEG 4000; condition 3: 0.2 M magnesium chloride, 0.1 M HEPES-Na pH 7.5, 30% (v/v) PEG 400; condition 4: 0.2 M calcium acetate, 0.1 M sodium cacodylate pH 6.5, 18% (w/v) PEG 8000. Crystals were observed on video images taken with an optical microscope (LEICA MZ APO) equipped with a color CCD camera (Sony).

Table 5. The effect of composition modification to the crystallization of β -amylase.

Drop volume ratio (protein:reservoir solutions)	1:2	1:1	2:1
Condition 5			

Crystallization conditions were chosen from Nextal Classics Suite 96 conditions. Condition 5: 0.2 M calcium chloride, 0.1 M HEPES-Na pH 7.5, 28% (v/v) PEG 400. Crystals were observed on video images taken with an optical microscope (LEICA MZ APO) equipped with a color CCD camera (Sony).

From these results we can see that the composition modification method may facilitate the attainment of the nucleation zone and improve the yield of crystals, in the cases where initial screening failures happened. Using composition modification can greatly increase the probability of reaching the nucleation zone to initiate crystallization, also the size, shape and quality of crystals can be improved. Moreover, by combining with the temperature screening methods, for example using phase diagram, more “hits” can be obtained in the trials of protein crystal growth.

3.2 Crystallogenesis study of API-A

3.2.1 Expression of API-A

In order to confirm the DNA sequence of the target gene in our experimental cells, *E. coli* BL21 (DE3), the DNA extraction experiment was carried out.

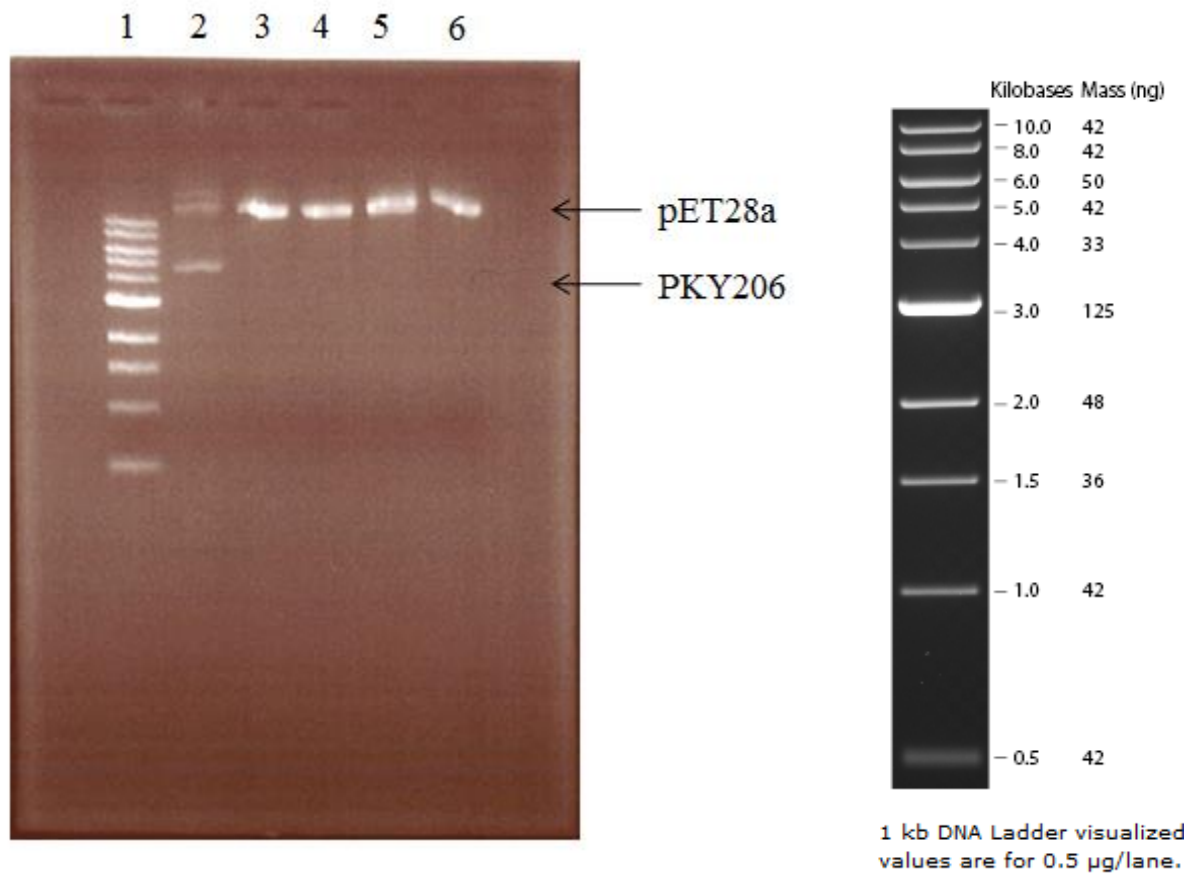


Figure 19. Agarose gel electrophoresis of DNA extracted from cultured cells. Lane 1: linear DNA ladder; lane 2: total plasmids in cell; lane 3-6: pET28a-derived expression vector containing the gene of API-A.

There are two antibiotic-resistance genes in the experimental cell: one encoding kanamycin-resistance in the pET28a-derived expression vector, and one encoding tetracycline-resistance genes in the PKY206. After the extraction of both plasmids from

experimental cells (Figure 19, lane 2), they were used to transform the DH 5 α competent cells. Then the transformed competent cells were cultured on agar plates, using 2 \times YT medium with 35 μ g/ml Kanamycin, in order to select the cells only containing plasmid of pET28a-derived expression vector with anti-Kanamycin gene, which expresses the target protein API-A (Figure 20, lane 3-6), for later DNA sequencing.

Table 6. Result of the pET28a-derived target gene sequencing. The sequence was obtained at the Sequencing and Genotyping Platform for CHUL Research Center. Start codon is indicated by red color, the codons of his tag are underlined (Annex: Full data of the pET28a-derived target gene sequencing).

GNNNNNNGNNNNAATTTCCCCTTCTNGAANAATTTTGTTTAACTTTAAGAAGGAGATA
TACCATGGGACACCATCACCATCACCATATGGATCCCGTCGTCGACAGCGATGGCGAT
GCGGTCCAGCTCAACTTGGGTGGCAACTACCCGCTATACACCATCCAGAGTGCTGCCA
TAGGCTTCCGCGGTGGGCTTTCCACATTGCGCAAGGACGCCTGCAAGAGCTACGTCT
ACGAGGCCCCCGAGACTGACCGCGGCTTGCCGGTGGGGTTCTCGGCATCGGCGACTT
CTCAGCCCGTCATGCAGCTGGGGTCCCGCTACAAGTTCTCCTTCTCGATGCCGGTACC
GTCATCTGCGACACCGCGTGGTCCATCGGCAAGTCGGAAACGAACGGTGGAAATCTC
CTTCCAGCCGATCACCGCCGGGGACTACTTTTACCTGAACAACCTTTAGCTGGTTCGAG
GCGAGGAGCACCGAGGAAACCGGCGTGTATAAGCTCGCTGCCTGCTCCTGTGAGTTC
TGCAAGATAGCTTGCCCCGAAGTAGGCTCCTTTAATGTCAACGGCCGTACCTTGCTGG
GCATCGGAGGGGAGCACTTCACCGTCCAGTTTCAGAAGTTCGACGCACTCTAAGCGG
CCGCACTCGAGCACCACCACCACCACCCTGAGATCCGGCTGCTAACAAAGCCCGA
AAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGG
GCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGATTGG
CGAATGGGACGCGCCCTGTAGCGGCGCATTAAAGCGCGGCGGGTGTGGTGGTTACGCG
CAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCT
TCCTTCTCGCCACGTTGCGCCGGCTTTCCCCGTCAAGCTCTAATCGGGGGCTCCCTTT
AGGGTTCCGATTTAGTGCTTTAC

These sequencing results (Table 6) show that, the gene coding API-A from our experimental cells was perfectly matching, to the API-A data published on PBD (Figure 9) (35).

3.2.2 Purification of API-A

The SDS-PAGE image of the proteins eluted from the nickel-nitrilotriacetic acid affinity resin (Ni-NTA) column is shown in Figure 20.

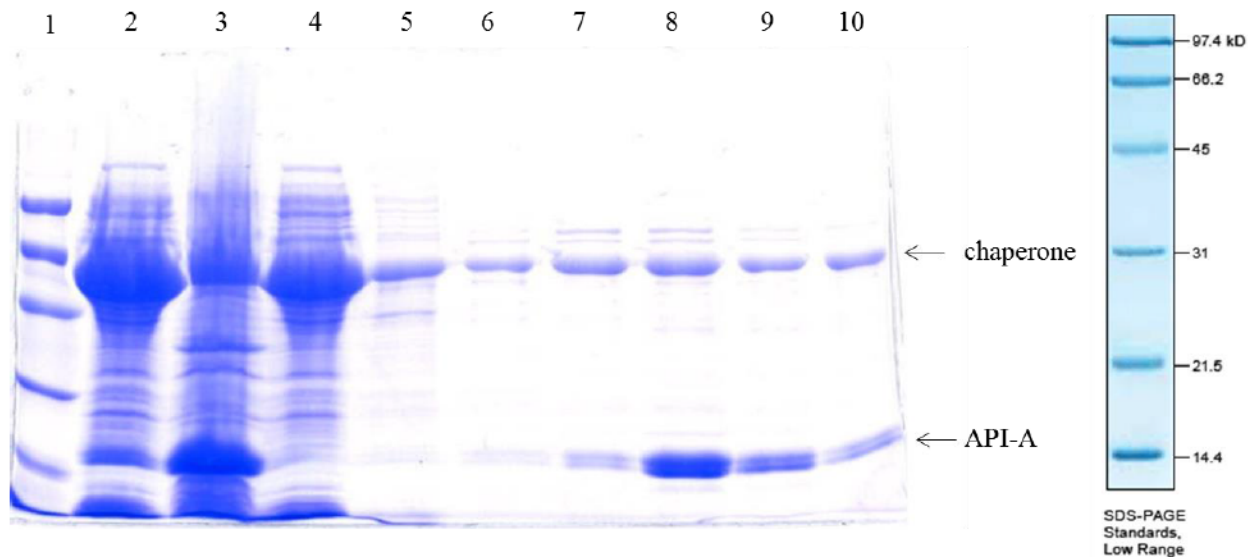


Figure 20. SDS-PAGE image of bacterial supernatant and Ni-NTA column fractions. Lane 1: molecular weight markers; lane 2: cell pellets; lane 3: bacterial lysate supernatant loaded onto the Ni column; lanes 4: column flow-through; lane 5: wash buffer A; lane 6: wash buffer B; lanes 7–10: fractions eluted with buffer containing 300 mM imidazole.

By comparing the API-A in lane 3 (supernatant from bacterial lysate) and lane 4 (column flow-through), it is very obvious that the His₆ tagged API-A efficiently bound to the Ni-NTA column, very few protein came out with the flow-through. The two washing steps removed most contaminant proteins and a small portion of API-A which did not strongly bind to the resin. The bound protein was then eluted by the lysis buffer plus 300 mM imidazole (Figure 20, lane 7-10). The eluted fractions mainly contained the chaperone protein and API-A, which made it much easier for later purification. Therefore, the Ni-NTA column was considered to be adequate for the first step of purification for API-A.

The SDS-PAGE image of the proteins eluted from the size exclusion chromatography column is shown in Figure 21.

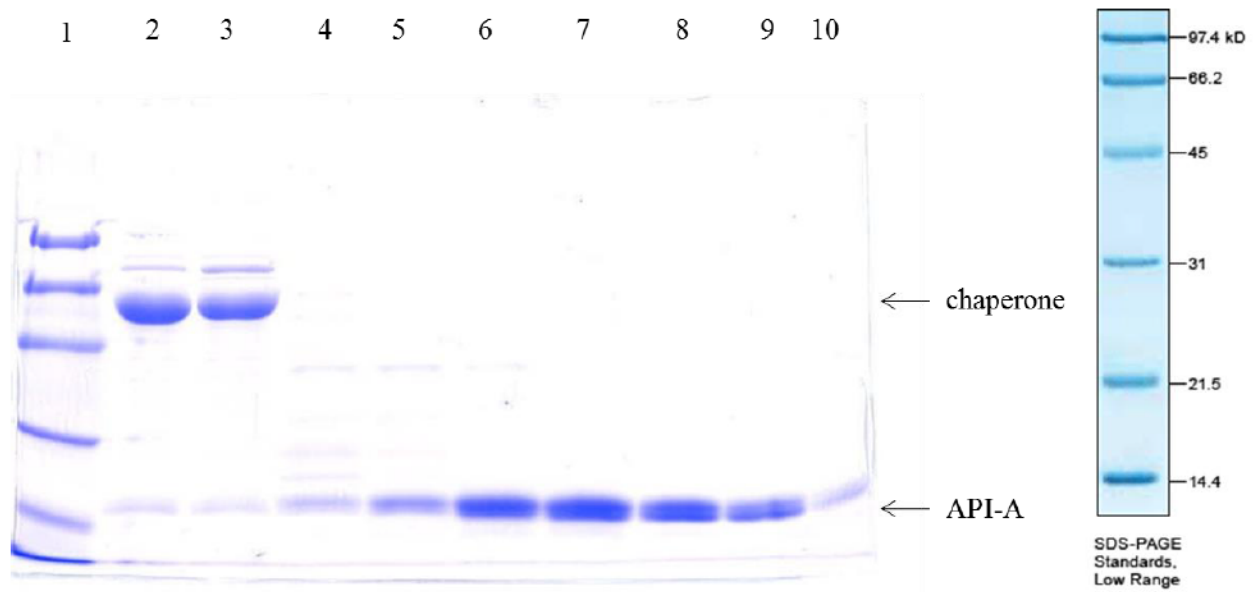


Figure 21. SDS-PAGE image of size exclusion chromatography fractions. Lane 1: molecular weight markers; lane 2, 3: Chaperone protein from plasmid PKY206; lane 4: unwanted proteins; lane 5-10: fractions of target protein API-A.

Moving along with the buffer, the protein of chaperone GroEL (Figure 21, lane 2, 3) of bigger molecular mass (about 60 kDa) was eluted first, then some contaminant proteins of the molecular mass between those of the chaperone GroEL and of API-A were eluted, which left API-A and GroES (Figure 21, lane 5-10) of molecular weight near 24 kDa and 10 kDa respectively to be finally eluted out. Considering the very close molecular mass of API-A and GroES, plus the band of GroES was too weak to be recognized in the SDS-PAGE, it is hard to know the possibility to separate the API-A from GroES by the Hiload 16/60 Superdex 200 column. Judging from the result, the size exclusion chromatography is also considered to be an adequate step for removing most contaminant proteins.

The SDS-PAGE image of the proteins eluted from the blue-sepharose affinity chromatography column is shown in Figure 22.

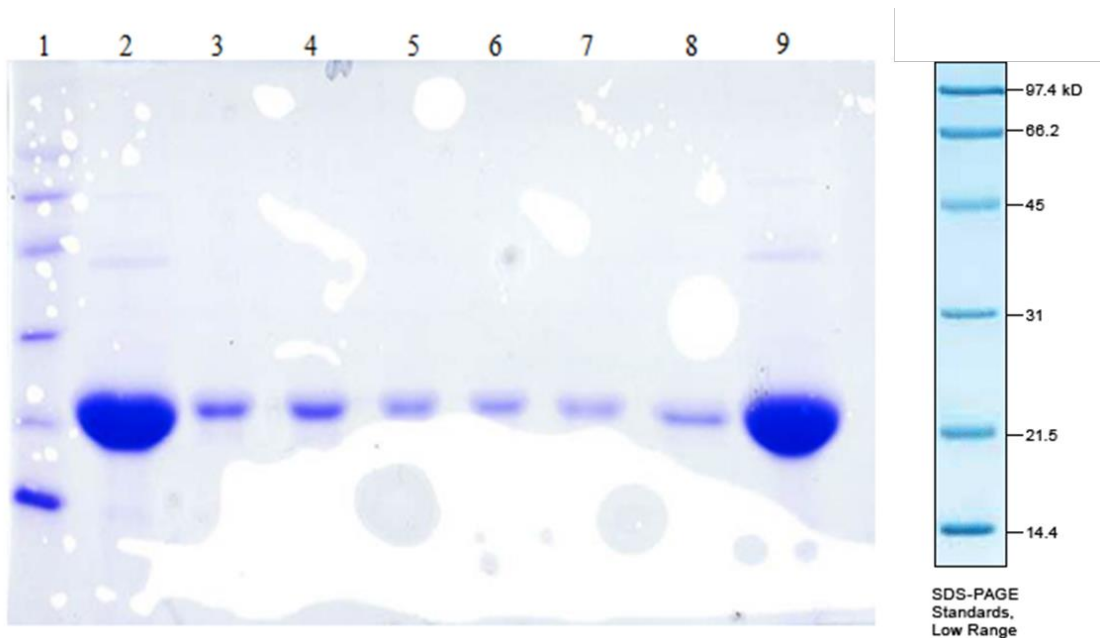


Figure 22. SDS-PAGE image of blue-sepharose affinity chromatography. Line 1: molecular weight markers; lane 2, 9: fractions from size exclusion chromatography; lane 3-7: upper band of API-A eluted with buffer A; lane 8: lower band of API-A eluted with 30% of gradient buffer B in the gradient buffer A.

From the fractions collected after size exclusion chromatography (Figure 22, lane 2, 9), it is obvious that the main bands were located at the molecular weight around 24 kDa, which indicates the purified product was mostly API-A. However, during the whole purification steps, the main bands of API-A contained two closely located bands which indicate that there are two proteins which are slightly different in molecular mass (Figure 20, Figure 21, Figure 22). By the means of mass spectrometry, the two bands were demonstrated to be both API-A (Figure 23). But they must be slightly different at their N- or their C-terminal ends.

After many crystallization trials, we concluded that the purified API-A from Ni-NTA column and size exclusion chromatography could not produce any crystals, which raised us a question about its quality, whether it was pure enough to be crystallized.

According to the requirement for the initial crystallization screening, the protein sample should be at least 90–95% pure on a Coomassie stained SDS-PAGE. The most important factor is that it must be conformationally homogenous (45, 65). Consequently the presence of two bands of

API-A that differ in molecular weight might be the problem that impedes its homogeneity. After many trials, the blue-sepharose affinity chromatography was found to be effective in separating the two bands of API-A. In the SDS-PAGE image (Figure 22), the upper band of API-A (Figure 22, lane 3-7) flowed through with the contaminant proteins such as GroES, whereas the lower band of API-A (Figure 22, lane 8) bound to the blue-sepharose column was then eluted using a linear salt gradient at 30% of gradient buffer B in the gradient buffer A.

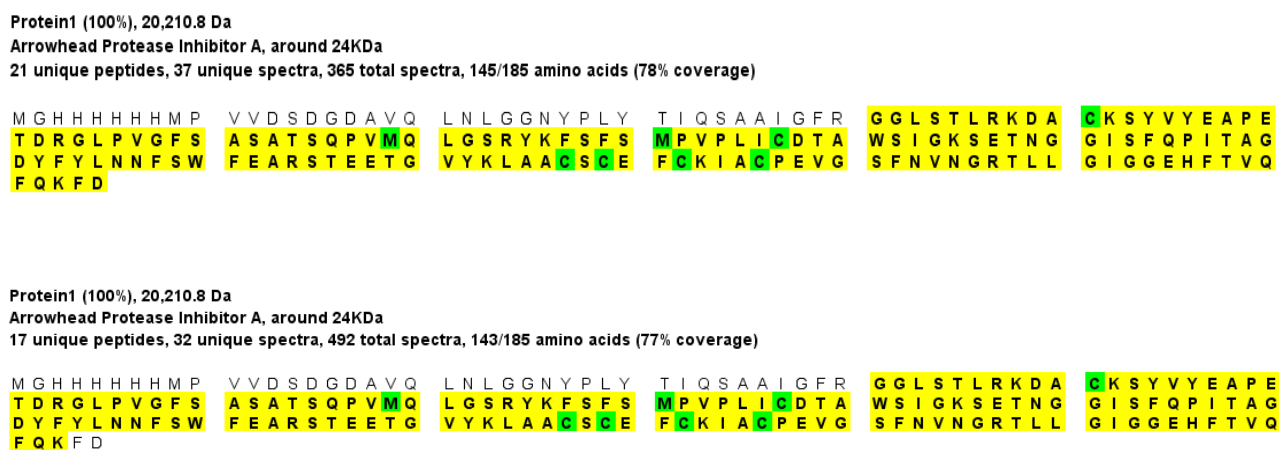


Figure 23. Peptide analysis of the two bands API-A (above is upper band and below is lower band) by the method of mass spectrometry.

In the peptide analysis by the method of mass spectrometry (Figure 23), the peptides of the two proteins match the sequence of API-A published in Protein Data Bank, which demonstrates that both proteins are API-A. The difference between the two proteins and the reason why they are different in apparent molecular mass are still unknown. It is assumed that at N- or C-terminal there are extra amino acids or they lack of some peptide caused by degradation.

3.2.3 Crystallization of API-A

After the three steps of chromatography mentioned above, the purified protein was concentrated to 4 mg/ml, the crystallization trials were carried out at 295 K by the hanging drop vapor

diffusion method using Nextal Classics Suite 96 conditions crystal screen kit. Each drop containing 1 μ l of reservoir solution and 1 μ l of protein sample was equilibrated against 0.5 ml of reservoir solution. Crystals appeared only in one condition, which is 0.2 M sodium acetate, 0.1 M TRIS-HCl at pH 8.5, 30% (w/v) polyethylene glycol 4000.

Then based on this crystallization condition, various methods were tried to improve the crystal quality. Relative crystallizability and composition modification were implemented by using the Nextal Classics Suite screen kit, however the quality of crystal didn't change much. Then different combinations of pH value, salt concentration and precipitant concentration were tried, good X-ray diffraction-quality crystals appeared at the condition of 0.2 M sodium acetate, 0.1 M TRIS-HCl at pH 8.3, 16% (w/v) polyethylene glycol 8000. 72 additive conditions were also applied to help the formation of a homogeneous, single crystal.

Finally, by using hanging drop vapor diffusion method at 295 K, at the crystallization condition of 0.2 M sodium acetate, 0.1 M TRIS-HCl at pH 8.3, 16% (w/v) polyethylene glycol 8000, within each drop there are 1 μ l of reservoir solution and 1 μ l of protein sample in addition of 0.25 μ l additive, the drop was equilibrated against 1 ml of reservoir solution, three main shapes of crystals were finally obtained, which were diamond, cubic and conic shaped (Figure 24).

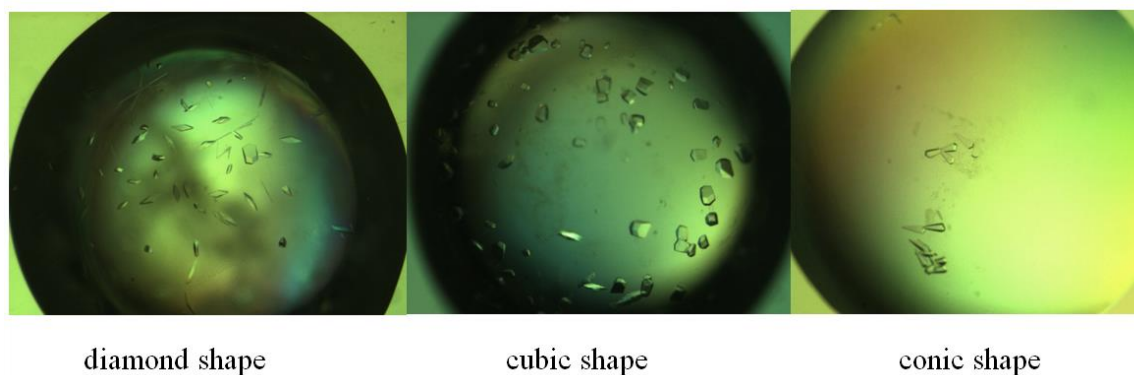


Figure 24. Main shapes of crystals obtained. The conic shape crystals were produced by adding jeffamine M-600 pH 7.0 (50%), the cubic shape crystals were produced by adding 1, 6-diamino-hexane (30%), whereas the diamond shape crystals were produced without additives.

3.3 X-ray data collection

The API-A crystals were sent to the Advanced Photon Source (APS) at Argonne National Laboratory for diffraction test. The diffraction pattern and the data collection statistics are shown below (Figure 25) (Tables 7 and 8).

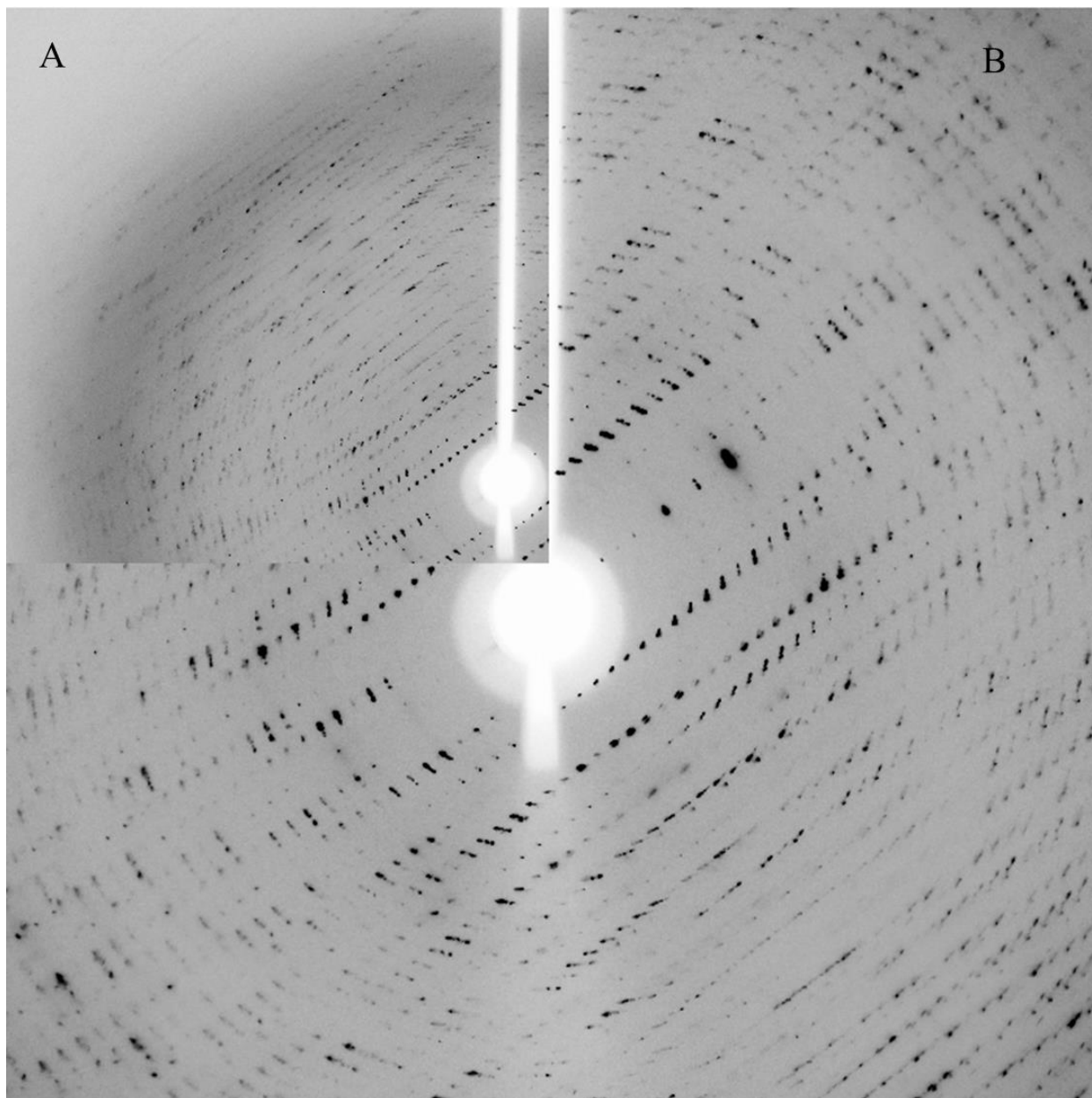


Figure 25. Diffraction pattern of API-A crystal. The part A shows the spots of high resolution area, and the part B shows the spots of low resolution area.

From Figure 25, which comes from a diamond shape crystal, numerous dots can be easily recognized, which formed the pattern of concentric circles. It is clear that this diffraction pattern was reflected by a protein crystal rather than an inorganic crystal, because the dots are located mainly in the center part of the film, and they form the shape of concentric circles. However, the diffraction spots are not well separated, spots in low resolution range are overlapped with each other, and spots in high resolution range are smeared. As it was reported that the cryo solutions or crystal handling can affect crystal quality (69), some future work can be done to get better data sets.

Table 7. The summary data collection.

Parameter	Data overall	Data outer shell
R _{merge}	0.297	0.411
R _{meas} (within I+/I-)	0.420	0.582
R _{meas} (all I+ & I-)	0.420	0.582
Fractional partial bias	-0.396	-0.388
Total number of observations	81237	11874
Total number unique	47974	7109
Mean(I)/sd(I)	3.1	2.0
Completeness	95.1	96.5

Table 8. The preliminary structural data from API-A crystal.

Parameter	Data
Cell dimensions	a=73.75 Å, b=73.75 Å, c=135.79 Å α=γ=90°, β=76.08°
Space group	P1
Resolution	3.10 Å

The cell dimensions of this crystal were: $a=73.75 \text{ \AA}$, $b=73.75 \text{ \AA}$, $c=135.79 \text{ \AA}$, $\alpha=\gamma=90^\circ$, $\beta=76.08^\circ$.

This crystal belongs to the space group P1 (Table 8).

The API-A structure in the API-A-trypsin complex was used as the search model in the molecular replacement trials. Solvent content analysis (Matthews coefficient) indicated that there are likely 8 molecules in the asymmetric unit in the apo API-A crystal.

Chapter 4

Conclusion

In this study, we used several crystallization methods, which are widely used in structural analysis of macromolecules, in order to get diffraction-quality crystals. All the methods are significantly important in dealing with different macromolecules with different physical and chemical properties. Methods like phase diagram and composition modification are very useful in facilitating the crystallization procedures; they can greatly increase the chance to succeed in crystallization trials, as well as to improve the crystal quality.

In the purification of proteins, many chromatography methods can be used based on their physical and chemical properties, but it often takes more than one method to attain homogeneity. Therefore a smart combination of different purification methods can be of great importance. In the purification of API-A, by analyzing its gene tag, the properties of the total proteins within the cultured cells, and also many experimental assays, the combination of nickel-nitrilotriacetate resin, size-exclusion chromatography and blue sepharose chromatography were used to get API-A of adequate conformational homogeneity for crystallization.

As important as the purification of protein, the crystallization of the target protein can be a bottleneck in obtaining the structural information. Choosing an adequate crystallization condition is not a simply logical work, but depends on numerous experimental trials. Some conditions which have the most probability to succeed in crystallizing proteins (personal communication from professor Sheng-Xiang Lin and Dr. Dao-Wei Zhu), were used as the first trials. The precipitation screens can be also very useful in finding a proper concentration of the target protein before carrying out the crystallization trials. The Nextal Classics Suite 96 conditions crystal screen kit was chosen to be the first trail in crystallization of API-A, based on which, further modification was realized in different combination of pH value, salt concentration and precipitant concentration as well as adding additives. Using hanging drop vapor diffusion method at 295 K, at the crystallization condition of 0.2 M sodium acetate, 0.1 M TRIS-HCl pH 8.3, 16% (w/v) polyethylene glycol 8000, crystals of apoenzyme API-A were obtained that diffracted X-rays at a resolution of up to 3.10 Å. Preliminary structure data were obtained: the unit cell dimensions are $a=73.75$ Å, $b=73.75$ Å, $c=135.79$ Å, $\alpha=\gamma=90^\circ$, $\beta=76.08^\circ$. This crystal

belongs to the space group P1.

Chapter 5

Perspectives

In the study of crystallogenesiis of API-A, there are two aspects that can be expected:

(1) The ideal size of a X-ray diffraction-quality crystal should be 0.1 mm in all dimensions; for the crystals of API-A, the cubic shape crystals have the right size but diffract at a very low resolution. On the other hand, the diamond shape crystals diffracted at higher resolution, but they do not reach the right size under the current conditions. Finally, in the present study, not many crystallization methods have been used with API-A. Accordingly, some more crystallization methods could be used to attempt improving both the size and the quality of apoenzyme API-A crystals. In addition, the cryo solutions or crystal handling could also be improved to help data collection.

(2) Once the adequate crystal of apoenzyme API-A will be obtained, the structure of apoenzyme API-A will be solved. It will be important not only to know the structure of this apoenzyme, but more interestingly, maybe some other possible active sites will be found, which can be helpful in finding other potential target proteases or some other possible substrate binding patterns. Also we could compare the API-A conformations both in its apo-protein status and in complex, for example the complex with two molecules of trypsin, to know its conformational changes when it binds to other molecules. This could contribute to the studies of protein-protein interactions, and also be meaningful in designing and developing other therapeutic inhibitors with similar catalytic mechanism.

Reference

1. Lopez-Otin, C., and Bond, J. S. (2008) Proteases: multifunctional enzymes in life and disease, *J Biol Chem* 283, 30433-30437.
2. Neurath, H., and Walsh, K. A. (1976) Role of proteolytic enzymes in biological regulation (a review), *Proc Natl Acad Sci U S A* 73, 3825-3832.
3. Rodriguez, D., Morrison, C. J., and Overall, C. M. Matrix metalloproteinases: what do they not do? New substrates and biological roles identified by murine models and proteomics, *Biochim Biophys Acta* 1803, 39-54.
4. Craik, C. S., Page, M. J., and Madison, E. L. Proteases as therapeutics, *Biochem J* 435, 1-16.
5. Turk, B. (2006) Targeting proteases: successes, failures and future prospects, *Nat Rev Drug Discov* 5, 785-799.
6. Schechter, I., and Berger, A. (1967) On the size of the active site in proteases. I. Papain, *Biochemical and Biophysical Research Communications* 27, 157-162.
7. Menard, R., Carmona, E., Plouffe, C., Bromme, D., Konishi, Y., Lefebvre, J., and Storer, A. C. (1993) The specificity of the S1' subsite of cysteine proteases, *FEBS Lett* 328, 107-110.
8. Neurath, H. (1984) Evolution of proteolytic enzymes, *Science* 224, 350-357.
9. Duffy, M. J., McGowan, P. M., and Gallagher, W. M. (2008) Cancer invasion and metastasis: changing views, *J Pathol* 214, 283-293.
10. Leung, D., Abbenante, G., and Fairlie, D. P. (2000) Protease inhibitors: current status and future prospects, *J Med Chem* 43, 305-341.
11. Garcia-Lorenzo, M., Sjodin, A., Jansson, S., and Funk, C. (2006) Protease gene families in Populus and Arabidopsis, *BMC Plant Biol* 6, 30.
12. Kraut, J. (1977) Serine proteases: structure and mechanism of catalysis, *Annu Rev Biochem* 46, 331-358.
13. Hedstrom, L. (2002) Serine protease mechanism and specificity, *Chem Rev* 102, 4501-4524.
14. Madala, P. K., Tyndall, J. D. A., Nall, T., and Fairlie, D. P. Update 1 of: Proteases Universally Recognize Beta Strands In Their Active Sites, *Chemical Reviews* 110, PR1-PR31.
15. Walker, B., and Lynas, J. F. (2001) Strategies for the inhibition of serine proteases, *Cell Mol Life Sci* 58, 596-624.
16. Blow, D. M. (1997) The tortuous story of Asp ... His ... Ser: structural analysis of alpha-chymotrypsin, *Trends Biochem Sci* 22, 405-408.
17. Dodson, G., and Wlodawer, A. (1998) Catalytic triads and their relatives, *Trends Biochem Sci* 23, 347-352.
18. Derewenda, Z. S., Derewenda, U., and Kobos, P. M. (1994) (His)C-H-O=C; Hydrogen Bond in the Active Sites of Serine Hydrolases, *Journal of Molecular Biology* 241, 83-93.
19. Matthews, D. A., Alden, R. A., Birktoft, J. J., Freer, S. T., and Kraut, J. (1975) X-ray crystallographic study of boronic acid adducts with subtilisin BPN' (Novo). A model for the catalytic transition state, *J Biol Chem* 250, 7120-7126.
20. Rawlings, N. D., Tolle, D. P., and Barrett, A. J. (2004) Evolutionary families of peptidase inhibitors, *Biochem J* 378, 705-716.
21. Bode, W., and Huber, R. (2000) Structural basis of the endoproteinase-protein inhibitor interaction,

- Biochim Biophys Acta* 1477, 241-252.
22. Kobayashi, H., Yagyu, T., Inagaki, K., Kondo, T., Suzuki, M., Kanayama, N., and Terao, T. (2004) Therapeutic efficacy of once-daily oral administration of a Kunitz-type protease inhibitor, bikunin, in a mouse model and in human cancer, *Cancer* 100, 869-877.
 23. Ishikura, H., Nishimura, S., Matsunami, M., Tsujiuchi, T., Ishiki, T., Sekiguchi, F., Naruse, M., Nakatani, T., Kamanaka, Y., and Kawabata, A. (2007) The proteinase inhibitor camostat mesilate suppresses pancreatic pain in rodents, *Life Sciences* 80, 1999-2004.
 24. Bunnage, M. E., Blagg, J., Steele, J., Owen, D. R., Allerton, C., McElroy, A. B., Miller, D., Ringer, T., Butcher, K., Beaumont, K., Evans, K., Gray, A. J., Holland, S. J., Feeder, N., Moore, R. S., and Brown, D. G. (2007) Discovery of Potent & Selective Inhibitors of Activated Thrombin-Activatable Fibrinolysis Inhibitor for the Treatment of Thrombosis, *Journal of Medicinal Chemistry* 50, 6095-6103.
 25. Ng, T. B., Lam, S. K., and Fong, W. P. (2003) A Homodimeric Sporamin-Type Trypsin Inhibitor with Antiproliferative, HIV Reverse Transcriptase-Inhibitory and Antifungal Activities from Wampee (Clausena lansium) Seeds, *Biological Chemistry* 384, 289-293.
 26. Bode, W., and Huber, R. (1992) Natural protein proteinase inhibitors and their interaction with proteinases, *European Journal of Biochemistry* 204, 433-451.
 27. Barrett, A. J., Rawlings, N. D., and Woessner, J. F. (2004) *Handbook of proteolytic enzymes*, 2nd ed., Elsevier Academic Press, Amsterdam ; Boston.
 28. Laskowski, M., and Kato, I. (1980) Protein Inhibitors of Proteinases, *Annual Review of Biochemistry* 49, 593-626.
 29. Nielsen, P. K., Bonsager, B. C., Fukuda, K., and Svensson, B. (2004) Barley alpha-amylase/subtilisin inhibitor: structure, biophysics and protein engineering, *Biochim Biophys Acta* 1696, 157-164.
 30. Song, H. K., and Suh, S. W. (1998) Kunitz-type soybean trypsin inhibitor revisited: refined structure of its complex with porcine trypsin reveals an insight into the interaction between a homologous inhibitor from *Erythrina caffra* and tissue-type plasminogen activator, *Journal of Molecular Biology* 275, 347-363.
 31. Vallee, F., Kadziola, A., Bourne, Y., Juy, M., Rodenburg, K. W., Svensson, B., and Haser, R. (1998) Barley alpha-amylase bound to its endogenous protein inhibitor BASI: crystal structure of the complex at 1.9 Å resolution, *Structure* 6, 649-659.
 32. Yang, H.-L., Luo, R.-S., Wang, L.-X., Zhu, D.-X., and Chi, C.-W. (1992) Primary Structure and Disulfide Bridge Location of Arrowhead Double-Headed Proteinase Inhibitors, *Journal of Biochemistry* 111, 537-545.
 33. Xie, Z.-W., Luo, M.-J., Xu, W.-F., and Chi, C.-W. (1997) Two Reactive Site Locations and Structure-function Study of the Arrowhead Proteinase Inhibitors, A and B, Using Mutagenesis, *Biochemistry* 36, 5846-5852.
 34. Yang, H. L., Wang, L. X., Zhu, D. X., and Qi, Z. W. (1991) Inhibitory property characterization and reactive site exploration of the arrowhead proteinase inhibitor, *Sci China B* 34, 832-839.
 35. Bao, R., Zhou, C. Z., Jiang, C., Lin, S. X., Chi, C. W., and Chen, Y. (2009) The ternary structure of the double-headed arrowhead protease inhibitor API-A complexed with two trypsins reveals a novel reactive site conformation, *J Biol Chem* 284, 26676-26684.
 36. Bode, W., and Huber, R. (1992) Natural protein proteinase inhibitors and their interaction with proteinases, *Eur J Biochem* 204, 433-451.
 37. Betts, M. J., and Sternberg, M. J. E. (1999) An analysis of conformational changes on protein-protein association: implications for predictive docking, *Protein Engineering* 12, 271-283.

38. Bradford, M. M. (1976) A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding, *Anal Biochem* 72, 248-254.
39. Domon, B., and Aebersold, R. (2006) Mass spectrometry and protein analysis, *Science* 312, 212-217.
40. Hernandez, P., Muller, M., and Appel, R. D. (2006) Automated protein identification by tandem mass spectrometry: issues and strategies, *Mass Spectrom Rev* 25, 235-254.
41. Vergara, A., Lorber, B., Zagari, A., and Giege, R. (2003) Physical aspects of protein crystal growth investigated with the Advanced Protein Crystallization Facility in reduced-gravity environments, *Acta Crystallogr D Biol Crystallogr* 59, 2-15.
42. Chayen, N. E. (1998) Comparative studies of protein crystallization by vapour-diffusion and microbatch techniques, *Acta Crystallogr D Biol Crystallogr* 54, 8-15.
43. Ducruix, A., and Giegé, R. (1999) *Crystallization of nucleic acids and proteins : a practical approach*, [2nd ed., Oxford University Press, Oxford ; New York.
44. Hampel, A., Labanauskas, M., Connors, P. G., Kirkegard, L., RajBhandary, U. L., Sigler, P. B., and Bock, R. M. (1968) Single crystals of transfer RNA from formylmethionine and phenylalanine transfer RNA's, *Science* 162, 1384-1387.
45. Giegé, R., and Mikol, V. (1989) Crystallogenesi of proteins, *Trends in Biotechnology* 7, 277-282.
46. Mikol, V., Rodeau, J. L., and Giege, R. (1990) Experimental determination of water equilibration rates in the hanging drop method of protein crystallization, *Anal Biochem* 186, 332-339.
47. Luft, J. R., Arakali, S. V., Kirisits, M. J., Kalenik, J., Wawrzak, I., Cody, V., Pangborn, W. A., and DeTitta, G. T. (1994) A macromolecular crystallization procedure employing diffusion cells of varying depths as reservoirs to tailor the time course of equilibration in hanging- and sitting-drop vapor-diffusion and microdialysis experiments, *Journal of Applied Crystallography* 27, 443-452.
48. McPherson, A. (1976) Crystallization of proteins from polyethylene glycol, *Journal of Biological Chemistry* 251, 6300-6303.
49. McPherson, A. (1995) Increasing the size of microcrystals by fine sampling of pH limits, *Journal of Applied Crystallography* 28, 362-365.
50. Rigaud, J.-L., Chami, M., Lambert, O., Levy, D., and Ranck, J.-L. (2000) Use of detergents in two-dimensional crystallization of membrane proteins, *Biochimica et Biophysica Acta (BBA) - Biomembranes* 1508, 112-128.
51. Jancarik, J., and Kim, S.-H. (1991) Sparse matrix sampling: a screening method for crystallization of proteins, *Journal of Applied Crystallography* 24, 409-411.
52. Guan, R.-J., Wang, M., Liu, X.-Q., and Wang, D.-C. (2001) Optimization of soluble protein crystallization with detergents, *Journal of Crystal Growth* 231, 273-279.
53. Taleb, M., Didierjean, C., Jelsch, C., Mangeot, J. P., Capelle, B., and Aubry, A. (1999) Crystallization of proteins under an external electric field, *Journal of Crystal Growth* 200, 575-582.
54. Wakayama, N. I., Ataka, M., and Abe, H. (1997) Effect of a magnetic field gradient on the crystallization of hen lysozyme, *Journal of Crystal Growth* 178, 653-656.
55. Suzuki, Y., Sasaki, G., Miyashita, S., Sawada, T., Tamura, K., and Komatsu, H. (2002) Protein crystallization under high pressure, *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology* 1595, 345-356.
56. Zhu, D. W., Garneau, A., Mazumdar, M., Zhou, M., Xu, G. J., and Lin, S. X. (2006) Attempts to rationalize protein crystallization using relative crystallizability, *Journal of Structural Biology* 154, 297-302.

57. Lin, Y.-B., Zhu, D.-W., Wang, T., Song, J., Zou, Y.-S., Zhang, Y.-L., and Lin, S.-X. (2008) An Extensive Study of Protein Phase Diagram Modification: Increasing Macromolecular Crystallizability by Temperature Screening†, *Crystal Growth & Design* 8, 4277-4283.
58. Zhang, C. Y., Mazumdar, M., Zhu, D. W., Yin, D. C., and Lin, S. X. (2011) Analysis and Statistics of Crystallisation Success Increase by Composition Modification of Protein and Precipitant Mixing Ratio, *Protein Pept Lett.*
59. Luft, J. R., Wolfley, J. R., Said, M. I., Nagel, R. M., Lauricella, A. M., Smith, J. L., Thayer, M. H., Veatch, C. K., Snell, E. H., Malkowski, M. G., and Detitta, G. T. (2007) Efficient optimization of crystallization conditions by manipulation of drop volume ratio and temperature, *Protein Sci* 16, 715-722.
60. Rayment, I. (2002) Small-scale batch crystallization of proteins revisited: an underutilized way to grow large protein crystals, *Structure* 10, 147-151.
61. Pflugrath, J. W. (2004) Macromolecular cryocrystallography--methods for cooling and mounting protein crystals at cryogenic temperatures, *Methods* 34, 415-423.
62. Blundell, T. L., Jhoti, H., and Abell, C. (2002) High-throughput crystallography for lead discovery in drug design, *Nat Rev Drug Discov* 1, 45-54.
63. Smyth, M. S., and Martin, J. H. (2000) x ray crystallography, *Mol Pathol* 53, 8-14.
64. Wuthrich, K. (1989) Protein structure determination in solution by nuclear magnetic resonance spectroscopy, *Science* 243, 45-50.
65. Drenth, J. (1999) *Principles of protein x-ray crystallography*, 2nd ed., Springer, New York.
66. Snyder, D. A., Chen, Y., Denissova, N. G., Acton, T., Aramini, J. M., Ciano, M., Karlin, R., Liu, J., Manor, P., Rajan, P. A., Rossi, P., Swapna, G. V., Xiao, R., Rost, B., Hunt, J., and Montelione, G. T. (2005) Comparisons of NMR spectral quality and success in crystallization demonstrate that NMR and X-ray crystallography are complementary methods for small protein structure determination, *J Am Chem Soc* 127, 16505-16511.
67. Brunger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T., and Warren, G. L. (1998) Crystallography & NMR system: A new software suite for macromolecular structure determination, *Acta Crystallogr D Biol Crystallogr* 54, 905-921.
68. Rosenbaum, D. F., and Zukoski, C. F. (1996) Protein interactions and crystallization, *Journal of Crystal Growth* 169, 752-758.
69. Baba, S., Hoshino, T., Ito, L., and Kumasaka, T. Humidity control and hydrophilic glue coating applied to mounted protein crystals improves X-ray diffraction experiments, *Acta Crystallogr D Biol Crystallogr* 69, 1839-1849.

Annex: Full data of the pET28a-derived target gene sequencing

>P11873_API-PET28A-1_T7P_A01_015.ab1

GNNNNNNGNNNNAATTTCCCCTTCTNGAANAATTTTGTTTAACTTTAAGAAGGAGATA
TACCATGGGACACCATCACCaT

CACCATATGGATCCCGTCGTCGACAGCGATGGCGATGCGGTCCAGCTCAACTTGGGTG
GCAACTACCCGCTATACACCAT

CCAGAGTGCTGCCATAGGCTTCCGCGGTGGGCTTTCCACatTGCGCAAGGACGCCTGC
AAGAGCTACGTCTACGAGGCC

CCGAGACTGACCGCGGCTTGCCGGTGGGGTTCTCGGCATCGGCGACTTCTCAGCCCG
TCATGCAGCTGGGGTCCCGCTAC

AAGTTCTCCTTCTCGATGCCGGTACCGCTCATCTGCGACACCGCGTGGTCCATCGGCA
AGTCGGAAACGAACGGTGAAT

CTCCTTCCAGCCGATCACCGCCGGGGACTACTTTTACCTGAACA ACTTTAGCTGGTTC
GAGGCGAGGAGCACCGAGGAAA

CCGGCGTGTATAAGCTCGCTGCCTGCTCCTGTGAGTTCTGCAAGATAGCTTGCCCCGA
AGTAGGCTCCTTTAATGTCAAC

GGCCGTACCTTGCTGGGCATCGGAGGGGAGCACTTCACCGTCCAGTTTCAGAAGTTC
GACGCACTCTAAGCGGCCGCACT

CGAGCACCACCACCACCACCTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAG
CTGAGTTGGCTGCTGCCACCGCTG

AGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCT
GAAAGGAGGAACTATATCCGGA

TTGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAAGCGCGGGGTGTGGTGGTTA
CGCGCAGCGTGACCGCTACACTT

GCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCCTTCTCGCCACGTTTCGC
CGGCTTTCCCGTCAAGCTCT

AATCGGGGGCTCCCTTTAGGGTTCCGATTTAGTGCTTTAC

>P11873_API-PET28A-2_T7P_B01_013.ab1

TNNNNNNNNATTTCCCCTCTAGAATAATTTTGTTTAACTTTAAGAAGGAGATATACCAT
GGGACACCATCACCATCACCA

TATGGATCCCGTCGTCGACAGCGATGGCGATGCGGTCCAGCTCAACTTGGGTGGCAA
CTACCCGCTATACACCATCCAGA

GTGCTGCCATAGGCTTCCGCGGTGGGCTTTCCACATTGCGCAAGGACGCCTGCAAGA
GCTACGTCTACGAGGCCCCCGAG
ACTGACCGCGGCTTGCCGGTGGGGTTCTCGGCATCGGCGACTTCTCAGCCCGTCATG
CAGCTGGGGTCCCGCTACAAGTT
CTCCTTCTCGATGCCGGTACCGCTCATCTGCGACACCGCGTGGTCCATCGGCAAGTCG
GAAACGAACGGTGGAAATCTCCT
TCCAGCCGATCACCGCCGGGGACTACTTTTACCTGAACAACCTTTAGCTGGTTCGAGGC
GAGGAGCACCGAGGAAACCGGC
GTGTATAAGCTCGCTGCCTGCTCCTGTGAGTTCTGCAAGATAGCTTGCCCCGAAGTAG
GCTCCTTTAATGTCAACGGCCG
TACCTTGCTGGGCATCGGAGGGGAGCACTTCACCGTCCAGTTTCAGAAGTTCGACGC
ACTCTAAGCGGCCGCACTCGAGC
ACCACCACCACCACCCTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAG
TTGGCTGCTGCCACCGCTGAGCAA
TAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAG
GAGGAACTATATCCGGATTGGC
GAATGGGACGCGCCCTGTAGCGGCGCATTAAAGCGCGGCGGGTGTGGTGGTTACGCGC
AGCGTGACCGCTACACTTGCCAG
CGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTCTCGCCACGTTCCGCCGGCT
TTCCCCGTCAAGCTCTAAATC
GGGGGCTCCCTTTAGGGTTCCGATTTAGTGCTTTACGGCA

>P11873_API-PET28A-3_T7P_C01_011.ab1

GGGGGNANAATTTCCCTCTAGGAATAATTTGTTTAACTTTAAGAAGGAGATATAC
CATGGGACACCATCACCaTCAC
CATATGGATCCCGTCGTCGACAGCGATGGCGATGCGGTCCAGCTCAACTTGGGTGGC
AACTACCCGCTATACACCATCCA
GAGTGCTGCCATAGGCTTCCGCGGTGGGCTTTCCACATTGCGCAAGGACGCCTGCAA
GAGCTACGTCTACGAGGCCCCCG
AGACTGACCGCGGCTTGCCGGTGGGGTTCTCGGCATCGGCGACTTCTCAGCCCGTCA
TGCAGCTGGGGTCCCGCTACAAG
TTCTCCTTCTCGATGCCGGTACCGCTCATCTGCGACACCGCGTGGTCCATCGGCAAGT
CGGAAACGAACGGTGGAAATCTC
CTTCCAGCCGATCACCGCCGGGGACTACTTTTACCTGAACAACCTTTAGCTGGTTCGAG
GCGAGGAGCACCGAGGAAACCG

GCGTGTATAAGCTCGCTGCCTGCTCCTGTGAGTTCTGCAAGATAGCTTGCCCCGAAGT
AGGCTCCTTTAATGTCAACGGC
CGTACCTTGCTGGGCATCGGAGGGGAGCACTTCACCGTCCAGTTTCAGAAGTTCGAC
GCACTCTAAGCGGCCGCACTCGA
GCACCACCACCACCACCCTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTG
AGTTGGCTGCTGCCACCGCTGAGC
ATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAA
AGGAGGAACTATATCCGGATTG
GCGAATGGGACGCGCCCTGTAGCGGCGCATTAAAGCGCGGCGGGTGTGGTGGTTACGC
GCAGCGTGACCGCTACACTTGCC
AGCGCCCTAGCGCCCGCTCCTTTTCGCTTTCTTCCCTTCCCTTCTCGCCACGTTCCGCCG
GCTTTCCCCGTCAAGCTCTAAA
TCGGGGGCTCCCTTTAGGGTTCGGATTTAGTGCTTTACGG

>P11873_API-PET28A-4_T7P_D01_009.ab1

NNNGNGNNNNATTTCCCTCTAGAATAATTTGTTTAACTTTAAGAAGGAGATATACCAT
GGGACACCATCACCATCACCA
TATGGATCCCGTCGTCGACAGCGATGGCGATGCGGTCCAGCTCAACTTGGGTGGCAA
CTACCCGCTATACCATCCAGA
GTGCTGCCATAGGCTTCCGCGGTGGGCTTCCACATTGCGCAAGGACGCCTGCAAGA
GCTACGTCTACGAGGCCCCCGAG
ACTGACCGCGGCTTGCCGGTGGGGTTCTCGGCATCGGCGACTTCTCAGCCCGTCATG
CAGCTGGGGTCCCGCTACAAGTT
CTCCTTCTCGATGCCGGTACCGCTCATCTGCGACACCGCGTGGTCCATCGGCAAGTCG
GAAACGAACGGTGGAAATCTCCT
TCCAGCCGATCACCGCCGGGGACTACTTTTACCTGAACAACCTTTAGCTGGTTCGAGGC
GAGGAGCACCGAGGAAACCGGC
GTGTATAAGCTCGCTGCCTGCTCCTGTGAGTTCTGCAAGATAGCTTGCCCCGAAGTAG
GCTCCTTTAATGTCAACGGCCG
TACCTTGCTGGGCATCGGAGGGGAGCACTTCACCGTCCAGTTTCAGAAGTTCGACGC
ACTCTAAGCGGCCGCACTCGAGC
ACCACCACCACCACCCTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAG
TTGGCTGCTGCCACCGCTGAGCAA
TAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAG
GAGGAACTATATCCGGATTGGC

GAATGGGACGCGCCCTGTAGCGGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGC
AGCGTGACCGCTACACTTGCCAG
CGCCCTAGCGCCCGCTCCTTTCGCTTTCCTCCCTTCCTTCTCGCCACGTTCGCCGGCT
TTCCCCGTCAAGCTCTAAATC
GGGGGCTCCCTTTAGGGTTCCGATTTAGTGCTTTACGGCA