Spring 4-27-2011

Conformational Bias in 2'-Selenium-Modified Nucleosides and the Effect on Helical Structure and Extracellular Recombinant Protein Production: Current Systems and Applications

Richard A. Thompson
rthompson32@student.gsu.edu

CONFORMATIONAL BIAS IN 2'-SELENIUM-MODIFIED NUCLEOSIDES AND THE EFFECT ON HELICAL STRUCTURE AND EXTRACELLULAR RECOMBINANT PROTEIN PRODUCTION: CURRENT SYSTEMS AND APPLICATIONS

by

RICHARD ADAM THOMPSON

Under the Direction of Dr. Markus W. Germann

ABSTRACT

Part One. X-ray crystallography has benefited from the synthetic introduction of selenium to different positions within nucleic acids by easing the solving of the phase problem. Interestingly, its addition to the 2' position of the ribose ring also significantly enhances crystal formation. This phenomenon was investigated to describe the effect of selenium-based and other 2' modifications to the ribose ring of nucleosides in solution, as well as the incorporation of the selenium-modified nucleotides into a helical structure. This work correlates the difference in conformation propensity between the selenium containing nucleosides and oligomers towards a rationale behind the enhanced crystal forming behavior. *Part Two.* Recombinant protein production is a critical tool in laboratories and industries, and inducing extracellular transport of these products to the culture medium shows potential for improving cases where the yields are not sufficient in quality or quantity. This review incorporates current practices and systems with future perspectives.

INDEX WORDS: Selenium-modified nucleic acids, NMR spectroscopy, Sugar puckering, Recombinant Protein Secretion

CONFORMATIONAL BIAS IN 2'-SELENIUM-MODIFIED NUCLEOSIDES AND THE EFFECT ON HELICAL STRUCTURE AND EXTRACELLULAR RECOMBINANT PROTEIN PRODUCTION: CURRENT SYSTEMS AND APPLICATIONS

by

RICHARD ADAM THOMPSON

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Science

in the College of Arts and Sciences

Georgia State University

2011

CONFORMATIONAL BIAS IN 2'-SELENIUM-MODIFIED NUCLEOSIDES AND THE EFFECT ON HELICAL STRUCTURE AND EXTRACELLULAR RECOMBINANT PROTEIN PRODUCTION: CURRENT SYSTEMS AND APPLICATIONS

by

RICHARD ADAM THOMPSON

Committee Chair: Dr. Markus W. Germann

Committee: Dr. Zhen Huang

Dr. Jenny Yang

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

May 2011

DEDICATION

To my Parents and Grandparents

ACKNOWLEDGEMENTS

Countless thanks to Dr. Germann for being accessible and accommodating to me throughout the pursuit of this degree, my fellow group members for their various assistance in my practical education, and to my family, friends, and colleagues who make up the support system I've relied on over the past years.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PART ONE

CONFORMATIONAL BIAS IN 2'-SELENIUM-MODIFIED NUCLEOSIDES AND THE EFFECT ON HELICAL STRUCTURE

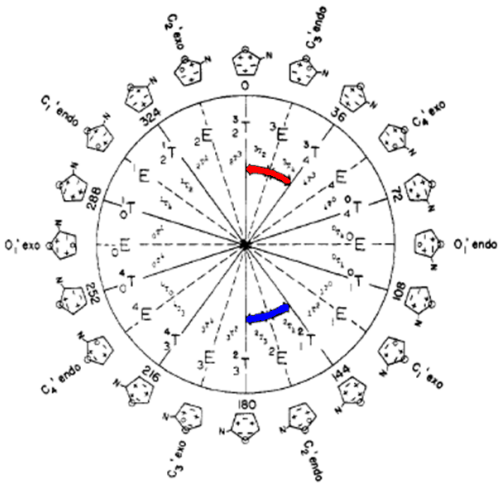
1 INTRODUCTION

1.1 Introduction to Nucleic Acid Structure

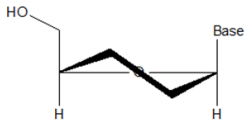
Nucleic acids are a vital component of any biological system and as such they are very widely studied. The cellular processes that nucleic acids and their constituents are a part of span from information or energy storage to catalysis or intercellular signaling. The driving force behind many nucleic acid functions, from drug or protein binding[4-6] to regulation of transcription or replication[7], is its structure. A monomer nucleotide is characterized by a nucleosidic base, a ribose sugar and a phosphate linker moiety, all of which can have an effect on the macroscopic structure of the molecule, but as the ribose units connect the base to the phosphate backbone, and being a five-membered ring prone to dynamic switching, or puckering, between conformations, the character of the sugar has a heavy influence on the overall configuration, especially in a double-stranded helix.

1.2 The Concept of Pseudorotation

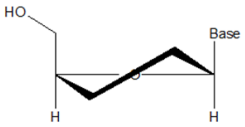
The characterization of this non-planarity of a ring system using the concept of pseudorotation was done on cyclopentane first by Kilpatrick et al.[8] and followed by others[9] who deduced the dynamic nature of these rings through various thermodynamic data. The concept was expanded by Altona and Sundaralingam who combined this concept with X-ray crystallography data to relate the five intracyclic torsion angles of a nucleosidic sugar in two pseudorotation parameters: phase angle (P) and puckering amplitude (Φ_m). They showed the rings essentially exist as two main types, North (3'-endo) and South (2'-endo), as designated by their phase angle[1] (Fig. 1.1). They further incorporated Karplus' relationship between torsion angle and H-H ^3J coupling



RNA	A-DNA	B-DNA	Z-DNA
--	Right-handed	Right-handed	Left-Handed
β-	β-	β-	B-
3'-endo	3'-endo	2'endo	2'endo, 3'endo
N-type	N-type	S-type	Both



3' endo



2' endo

Figure 1.1 Pseudorotation Wheel and Sugar Puckering Conventions (Left)Pseudorotation wheel depicting the conformation designated by a given phase angle, P. The regions shaded designate the range northern (red) and southern (blue) sugars populate in the wheel. Image adapted from Altona and Sundaralingam (1972). (Right) Trends in nucleic acid structures, decreasing: helical handedness, glycosidic base orientation, sugar pucker designation, pseudorotation phase angle designation.[1-3] (Bottom Right) Visualization of the two generalizd conformers discussed.

values[10-11] to describe the two state equilibrium that furanose rings exhibit and through various refinements have allowed for quantification of the percentage of either form that exists at equilibrium.[12] No experimental data has suggested the use of a third pseudorotation parameter, and it has since also been shown that different forms of nucleic polymers prefer different sugar conformations.[13] A brief overview of trends in different nucleic acid structures is given in Figure 1.1. The Altona-Sundaralingam (AS) formalism is a powerful tool for extracting structural information based on coupling constants, which will be employed later.

1.3 Nucleic Acids in X-ray Crystallography

Since many techniques are used to investigate structural properties of nucleic acids, being able to reconcile the disadvantages of these techniques is beneficial to

develop a clear understanding of molecular processes. X-ray crystallography has been used to much success to study structural characteristics of many biological macromolecules, but to overcome the difficulties of crystallization and phase determination force crystallographers to alter the natural structure of these macromolecules with heavy-atom soakings or modifications. An additional problem in this case is that methods that are very effective with proteins have been proven to be much more difficult in DNA and RNA.[14] For instance, bromine derivatization can be problematic because it acts as a good leaving group and can attract nucleophilic attack if it is positioned anywhere on the furanose ring. Bromine addition to the base can lead to decomposition when exposed to UV light, as exhibited in photo-crosslinking of nucleic acids to proteins in order to determine contact points.[15]

1.4 Selenium Modifications in Biopolymers

Interestingly, the incorporation of selenium atoms into a macromolecule, which has been shown to work well in protein crystal samples,[16] has recently been explored by Huang et al. as a method for DNA or RNA structural investigation.[17-19] They have reported minimal disruption of structure between crystal structures of unmodified DNA oligomers and derivatives with 2'-selenomethyl and 5-bromine modifications, and the sugar pucker of all of these molecules are found to be A-form DNA, having a 3'-endo conformation.[14] More remarkably, they also report a much quicker rate of crystal formation with the selenomethyl modification than the bromine derivative or the unmodified control,[20] which raises questions about the effect the methyl-selenium modification has on crystal stability or desolvation rates considering the sugar pucker is the same in a crystal whether the 2' position is modified or not. It is also worth noting

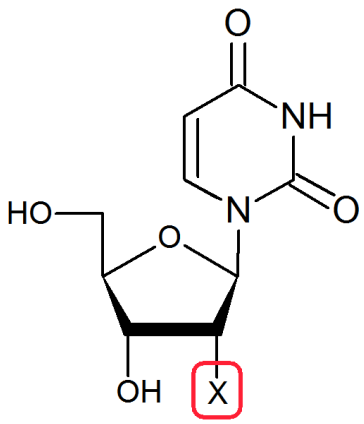


Figure 1.2 Modified Nucleosides used in this study. X= H (deoxyuridine, compound **1**); OH (uridine, **2**); OCH_3 (**3**); F (**4**); SCH_3 (**5**); $SeCH_3$ (**6**).

that in solution-based biophysical studies of nucleic acids, a 2' substituent is a determinant of the sugar conformation and its dynamics,[21-22] so its effect on crystal formation could be a result of this.

1.5 Goals of This Study

This work looks at the 2'-methylseleno derivative of uridine free in solution to determine its propensity towards one conformation or the other and to compare this behavior to that of other 2'-uridine substitutions. It also addresses the structural origin for the facilitated crystal formation by investigating the duplex structures containing selenium modifications using NMR, melting temperatures, ethidium bromide fluorescence and molecular modeling. The nucleoside behavior is then compared to the crystal structures and other data of the selenium modified oligomers. The library of 2' substituted uridines is described in Figure 1.2, while the sequences used in the oligomer studies are presented in Table 1.1.

Table 1.1 Duplex sequences for NMR and T_M studies, U^{Se} is compound **6**

I	5'-d(CATGCATG)
II	5'-d(GCGAATTCGC)
III	5'-d(GCGAAU^{Se}TCGC)
IV	5'-d(CGCGAATTCGCG)

2 MATERIALS AND METHODS

2.1 Nucleoside Studies

The sulfur and selenium based modifications were prepared as reported,[14,23] and all other compounds were purchased from Tech Chem. Nucleoside experiments were performed with a Bruker Avance 500 MHz spectrometer equipped with a TBI triple-resonance broadband capable probe head at 298K. Samples were prepared to be 1.0 mM nucleoside in D_2O with 10 mM sodium phosphate adjusted to pH* 6.0. DSS was used as an internal standard. Routine 1D ^{1}H NMR experiments with water presaturation pulses were performed on each nucleoside in order to confirm purity of the samples and to measure coupling constants. Double quantum filter COSY experiments (32 scans) were recorded to confirm assignments. A low-flip angle COSY was recorded for deoxyuridine (**1**) to clarify couplings caused by the 2' and 2'' protons.

2.2 Computational Parameters

DAISYSIM, a component of Topspin 2.1 (Bruker), was used to simulate spectra from the acquired NMR data in order to precisely determine the individual couplings and chemical shifts. DAISYSIM refines coupling and chemical shift estimates by a user-directed iteration algorithm. The refined coupling constants were used as the input into PSEUROT 6.0[24] to calculate the pseudorotation parameters according to established

practices. Also, in an attempt to move from a command line style of input to a more modern, user-friendly, GUI-based computational method, a Matlab-based (Mathworks) pseudorotation program was used for further substantiation.[25] PSEUROT 6.0 has been used to much success to calculate pseudorotation parameters of pentose rings from NMR data in several instances[26-28] and was provided by Altona and de Leeuw. The Matlab program was provided through a GNU General Public License by Hendrickx and Martins. The computation for each compound was initially set up with the conditions described in the user's manual of PSEUROT 6.0. The initial %S conformer was varied in subsequent trials in order to alleviate any bias built into the program with respect to conformational preference. Each of these initial states was refined during the computation by each program to give a theoretical pure N- and S-conformer population which was used to fit the data. The change in electronegativity of the 2' substitutions was accounted for in the input file; the values are derived from a Huggin's based electronegativity scale referenced to hydrogen specifically for use with generalized Haasnoot-Karplus equation as suggested by the authors of PSEUROT 6.0.[29-32] The Matlab program, since it was designed with the same computational premises, suggested the same values in the User's Manual.[25] The input and output data from each program are compiled in the Appendices.

2.3 Melting Temperature Assays

The melting assay was performed on the control and modified duplexes (**II** and **III**, respectively), through absorbance monitoring at 274 nm. The buffer was prepared to 400 mM sodium chloride, 10 mM sodium phosphate, and 0.1 mM EDTA at pH 6.5. The concentration of the both strands was set to 8 μM. Also, a second selenium sample was

prepared in the same buffer with 32 μM DNA in order to quantify the effect of concentration on the formation of a duplex. During the T_M assay, the temperature was reversibly ramped from 20°C to 90°C at 0.3°C/ min, controlled by a Cary spectrometer and heating block.

2.4 Ethidium Bromide Fluorescence

Oligomer samples of increasing length (octamer: **I**, decamer: **II**, Se-decamer: **III**, dodecamer: **IV**) were 15 μM in nucleotides or ~0.8 μM in duplex concentration and contained 1 μg/mL ethidium bromide, 100 mM NaCl, 10 mM sodium phosphate and 0.1 mM EDTA at pH 6.5. Samples were individually placed into PCR tubes and imaged on a Typhoon 9400 Variable Mode Imager from Amersham Biosciences. Excitation for imaging occurred at 532 nm and emission was measured at 610 nm.

2.5 Imino Proton Observation

NMR samples of sequences **II** and **III** were prepared at 50mM sodium chloride, 10 mM sodium phosphate, 0.1 mM EDTA and pH 6.4 in 9:1 H_2O:D_2O. Imino proton spectra were recorded on an Avance 600 MHz spectrometer using jump and return water suppression according to established practice.[33] Selenium samples (sequence **III**) were prepared at strand concentrations of 100 and 20 μM (designated high and low, respectively). The control sequence was prepared at 250 μM, in order to minimize acquisition time.

2.6 Model Development

Standard A- and B-form DNA helical models of sequence **III** were built within Spartan06 (Wavefunction) to estimate the position of 2'-$SeCH_3$-modification inside each

secondary structure. The A-form set the sugar puckering as N–type and with a rise and twist of 2.548 Å and 32.7^{o} per base, respectively, as described in the Spartan manual. The second model was made to be B-form (S-type, 3.375 Å, 36^{o}). After the models were built, the modification was inserted into the 2' position.

3 RESULTS

3.1 Data Fitting and NMR Assignments

In order to obtain an accurate description of the relevant coupling constants from the NMR data, the spectra were simulated from within TOPSPIN 2.1. In order for the simulation to be properly designed, the proper assignment of residues is critical. 2-D COSY spectra for each compound in Figure 1.2 were obtained to fully assign the peaks with high confidence. The correlation of resonances seen in the spectra was used to fully assign the sugar ring protons. The diagonal peaks arise from the peaks in each dimension seeing themselves (i.e. y = x), and from the diagonal one can determine which other peaks are within 3 bonds of the peak of interest. Knowing that the 1' proton should only see one resonance, one can follow the rest of the correlation pathway around the ring. Figure 3.1 shows the COSY spectrum of compound **6**, with the pathway highlighted. This strategy was repeated for each compound in this study and the full assignment of the ring protons was determined. Using the assignments, the inputs for the simulations were created.

Using DAISYSIM, a spin system simulation was fit to the NMR data according to a qualitative assessment by the user, i.e. if the simulation has not been accurately laid over the actual data, then further simulations and refinements are made in a recursive

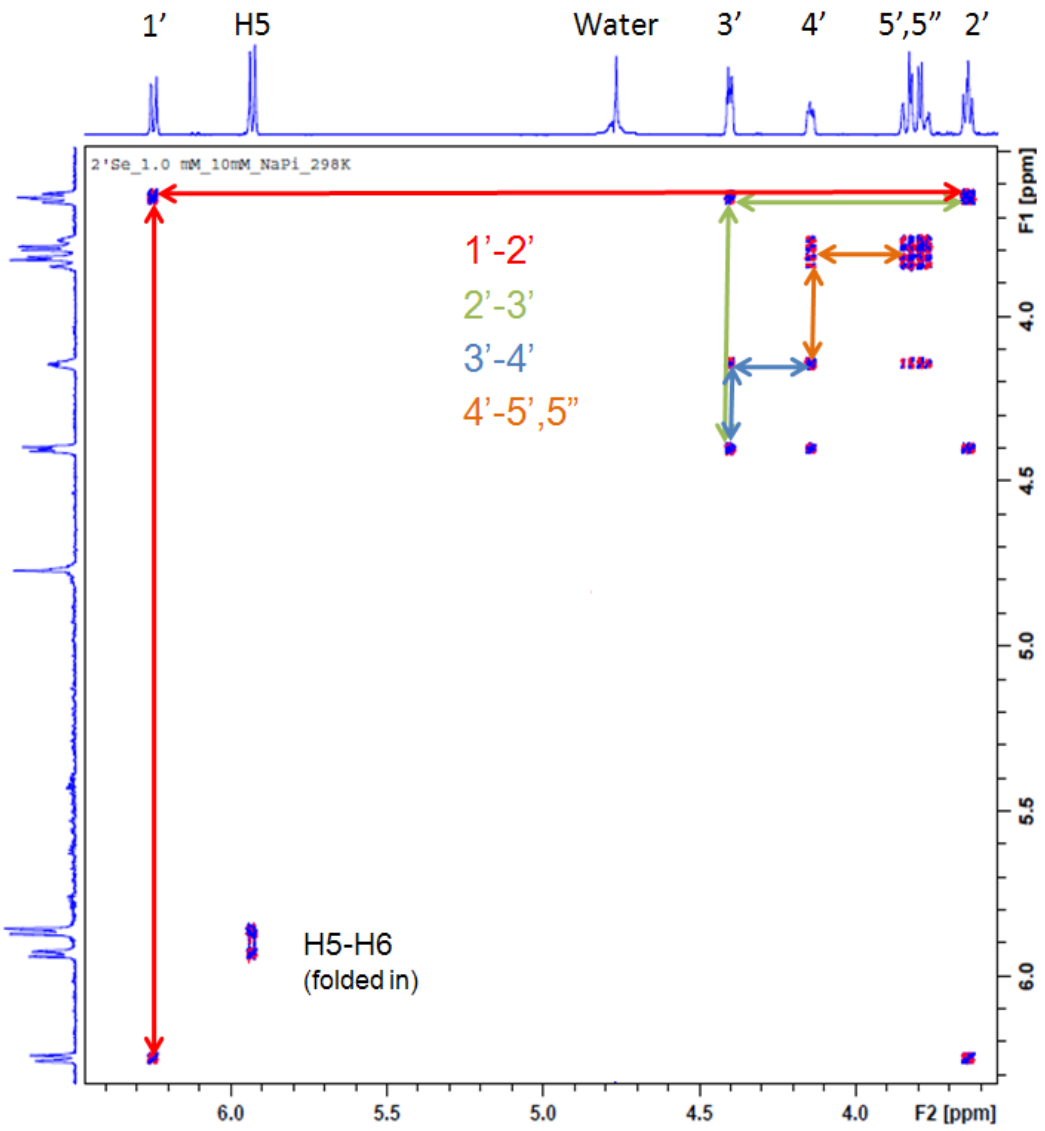


Figure 3.1 COSY spectrum of 2'-methylseleno-uridine, compound **6.** Assignment pathways are colored.

fashion until the simulation fits the data appropriately. This trial and error type method worked well in this situation but might not be the most effective way to determine obscure coupling constants from complicated systems. Nevertheless, this method was able to simulate the data to a high level of accuracy, although there was not an RMSD value returned by the fitting program, the experimental data and the simulation corresponded to each other fluently. Figure 3.2 shows the data fitting in the

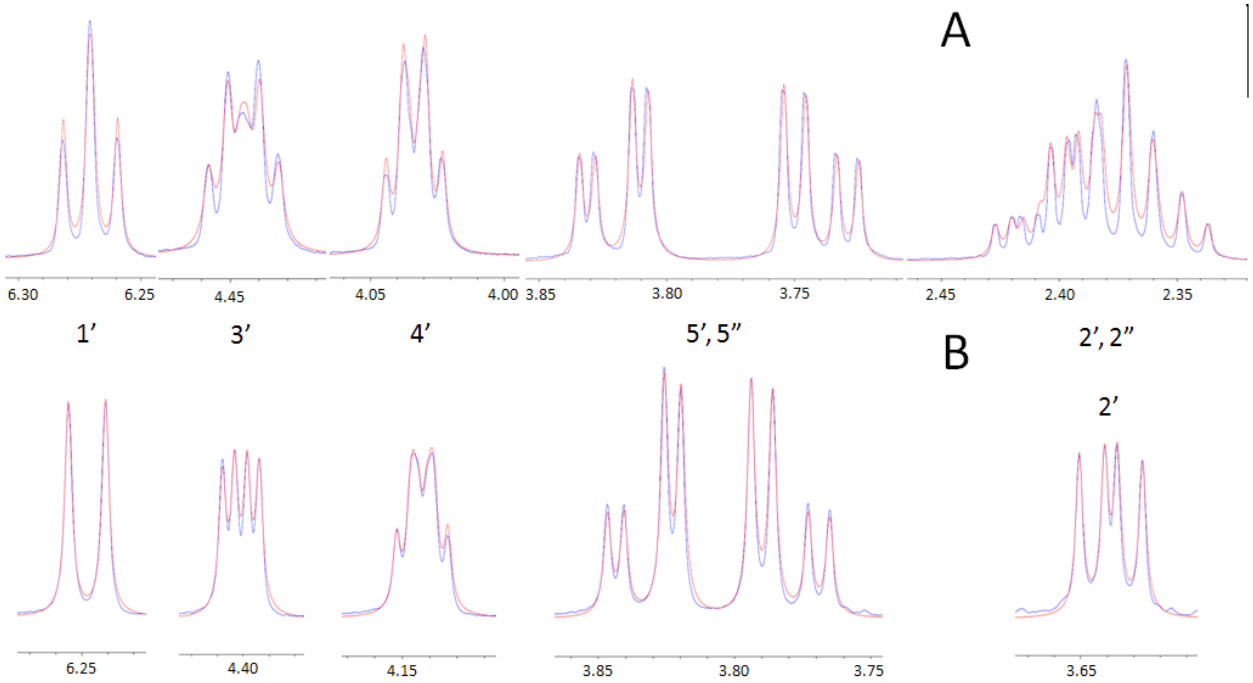


Figure 3.2 Simulation vs. NMR Data. (A) Deoxyuridine, compound **1**. (B) 2'-selenomethyl-deoxyuridine, compound **6**. Blue line is NMR data, red line is simulation results.

1-D ^{1}H spectra of compounds **1** and **6**. Upon first looking at the spectra, without considering the COSY spectra, the splitting patterns make sense when comparing the two. The 1' proton is split by two signals in the spectrum of compound **1**, corresponding to the 2' and 2" protons. In compound **6**'s spectrum, the 1' peak is only split by one proton, at the 2' position, because the 2" proton has been replaced by the methylseleno group in compound **6**. The 3' signal in **1**'s spectrum is split by an extra signal as well, as evident when comparing to **6**, following the same logic. This observation combined with the COSY spectra gives a high level of confidence in the data obtained from the simulations.

3.2 Nucleoside Characterization via Pseudorotation Calculations

The data from the NMR experiments is the backbone of the subsequent study, and all of the coupling values obtained by fitting the raw data in DAISYSIM were used in the pseudorotational calculations. The experimental data is compiled in Table 3.1 and shows reasonable correlation with literature reports. The trends in ^{3}J values and chemical shift with respect to substituent identity begin to reveal themselves even before pseudorotation parameters are calculated, and hint at the behavior of the sugar ring and the effects of the different modifications. Specifically, the $^{3}J_{1'-2'}$ and $^{3}J_{3'-4'}$ values, which arguably are the most affected by a change from 2'-endo to 3'-endo conformations, are the most dynamic of the data collected and sets the stage for explanation through a pseudorotation perspective. The fitted coupling data was used as the input parameters for the calculation of pseudorotation values as described in Section 2.2. Since all endocyclic coupling constants were known, the discrepancies arising from the mathematical determination of pseudorotation parameters, i.e. five torsion angle expressions with five variables, were minimized by eliminating solutions which did not fit within the whole set of equations. The optimized conformations and the percent of each were similar between PSEUROT and Matla and correlated with published results.[23,34-37] In Table 3.1, the top portion tabulates the NMR data, while the bottom portion shows the output from PSEUROT 6.0 (PS) and the Matlab program (ML). The pseudorotation data is also compiled in Table 3.1. It is interesting to see how the two starting conformations do not differ much between compounds, (the range is roughly 50^{o}) but the percent S conformation varies significantly. The data is believable because it follows literature reporting and implies that an increase in substituent electro-

Table 3.1 Compiled NMR and Pseudorotation Data

X	H			OH			OCH_3			F*			SCH_3			$SeCH_3$	
	Exp		Lit[35]	Exp		Lit[35]	Exp		Lit[34]	Exp		Lit[35]	Exp		Lit[23]	Exp	
J1'-2'	7.2		6.3	4.5		4.2	3.9		3.6	1.4		1.5	8.3		8.5	8.7	
J 1'-2''	6.1		6.4							19.7		19.7					
J 2'-3'	6.9		6.3	5.3		5.3	5.2			5.0		5.1	5.8		5.5	5.7	
J2''-3'	3.9		4.3							21.5		21.6					
J 3'-4'	3.9		4.0	5.5		5.7	5.7			8.6		8.7	2.8		2.0	2.9	
J H5-H6	8.1		8.1	8.1		8.0	8.1		8.3	8.1		8.1	8.1			8.1	
δ2'	2.4		2.4	4.3		4.3	4.1			5.2		5.2	3.6		3.4	3.6	
δ4'	4.0		4.0	4.1		4.1	4.1			4.2		4.1	4.2		3.9	4.1	
	PS	ML	Lit[35]	PS	ML	Lit [35]	PS	ML	Lit [34]	PS	ML	Lit [35,36]	PS	ML	Lit [23,37]	PS	ML
I.																	
N Type																	
P	18.0	2.3	18	32.4	13.5	18	12.5	34.4	11	28.6	36.1	21	-22.5	50.6	--	-13.3	7.2
Φ_M	38.0	33.5	--	32.0	30.5	--	32.0	35.5	35	32.0	34.2	--	32.0	16.7	--	32	20.3
II.																	
S type																	
P	141.5	149.7	162	156.6	129.40	162	144.7	162.7	171	38.6	-4.0	159	138.7	127.1	--	137.6	134.7
Φ_M	32.3	22.9	--	35.0	41.70	--	35.0	30.9	37	35.0	34.2	--	35.0	45.1	--	35.0	40.4
%S	0.67	0.58	0.6	0.40	0.45	0.4	0.38	0.39	0.4	0.32	0.03	0.17	0.81	0.76	0.76	0.84	0.83

negativity will drive the system into a state with a higher %N conformation,[38] as **1** favors a predominately S mixture while **2**, **3**, and **4** prefer an increasingly N mixture. Compound **4**, despite varying the starting conditions with different starting mixtures, converged to 'equilibrium' where both conformers were of N-type, essentially implying that the S-type conformer does not exist free in solution under these conditions. Both PSEUROT and Matlab returned this output. The literature shows that chlorine and bromine substitutions fit the overall relationship between electronegativity and percent S; the report of a 50-50 mixture of conformers makes sense[36] as these atoms have an electronegativity value less than oxygen but more than hydrogen. This trend is no longer observed, however, when considering the sulfur and selenium based compounds. Compounds **5** and **6** are found to more strongly prefer the S conformation than compound **1** in solution, which is the opposite of what the electronegativity or crystal structures suggest. Since the programs correlate well with literature results, a computational error is unlikely and an inference can be made that steric effects between the 2' substituents and the base drive the preference of the S conformation. There is strong correlation between various NMR data points and the %S value, which is an intrinsic principle of the programs themselves, but suggests that reasonable prediction of sugar puckering dynamics can be made from raw NMR data. As stated above, the $^{3}J_{1'-2'}$ and $^{3}J_{3'-4'}$ values are the most dynamic because they are the most affected by a change from 2'-endo to 3'-endo conformations. Especially relevant is the $^{3}J_{3'-4'}$ couplings because they are affected by the ring dynamics, and would be only minimally impacted by the 2'-substituent identity and when plotted against %S, as in Figure 3.3, show linear behavior. This expands on the graphical method presented by Rinkel and Altona,[39]

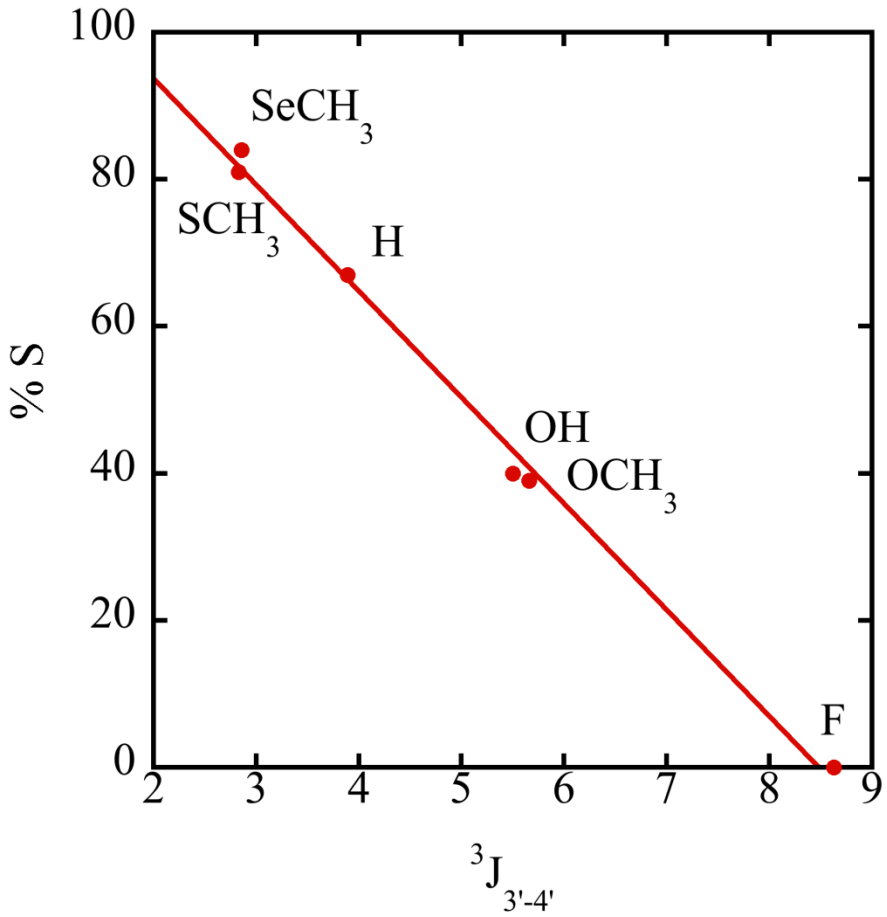


Figure 3.3 Relationship between $^{3}J_{3'\text{-}4'}$ coupling constant (Hz) and %S conformation

and works when individual couplings are known. These results clearly establish that 2'-$SeCH_3$-modified nucleosides strongly prefer a 2'-endo conformation in solution. However, this is exactly the opposite of what is observed in the crystal structures. This discrepancy was an interesting revelation and prompted further investigation of selenium-containing nucleosides within a duplex in solution.

3.3 Duplex Stability

In order to gain perspective on the physical effects of the 2'-$SeCH_3$-modification to a double stranded DNA molecule in solution, fluorescence, T_M and NMR data were

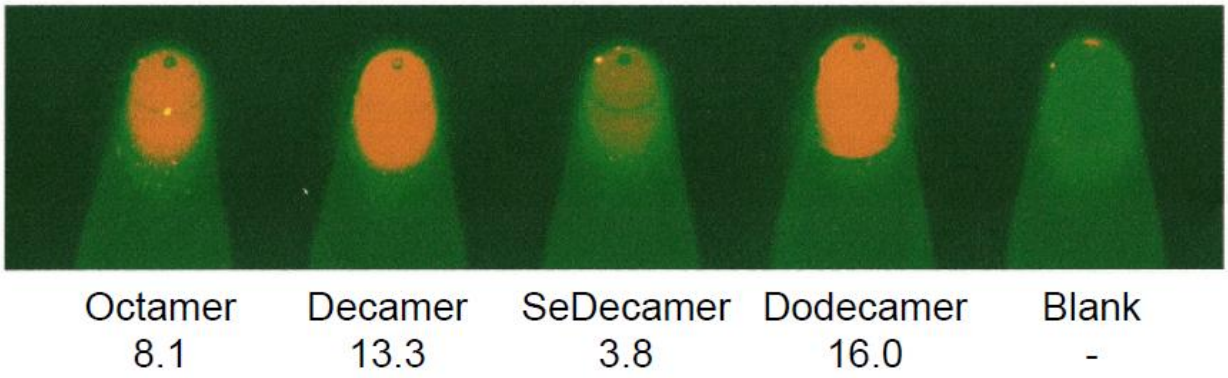


Figure 3.4 Duplex stability from ethidium bromide fluorescence. DNA samples (Octamer: **I**, Decamer: **II**, SeDecamer: **III**, Dodecamer **VI**) containing contain 1µg/mL ethidium bromide were placed in PCR tubes and imaged (Excitation at 532 nm, emmision at 610 nm). Relative fluorescence data, corrected for the blank, is indicated for each sample.

acquired and examined. The results of the ethidium bromide intercalation study (Figure 3.4) showed a clear discrepancy in the fluorescence of the different samples. The fluorescence intensity typically increases as a function of oligonucleotide length as more intercalation sites are possible in longer sequences. The unmodified octamer, decamer, and dodecamer exhibited this behavior. However, the Se-decamer displayed a lower fluorescence than the octamer, alluding to the destabilization effects of the modification.

The difference in melting curves between the unmodified decamer and the selenium decamer is also immediately noticed. The stability of the self complementary sequence **III** containing one 2'-$SeCH_3$-modification and its control **II** was determined by UV melting (Figure 3.5). The control duplex forms a standard B-type helical structure and exhibits a regular melting profile with an expected stability.[40-41] On the other hand, the shifted and shallow melting curve for the DNA strand **III** containing a single 2'-$SeCH_3$-modification demonstrates through a change in hyperchromicity that duplex formation was seriously destabilized. The observation that the denaturation was not concentration dependent also suggests the involvement of intramolecularly formed

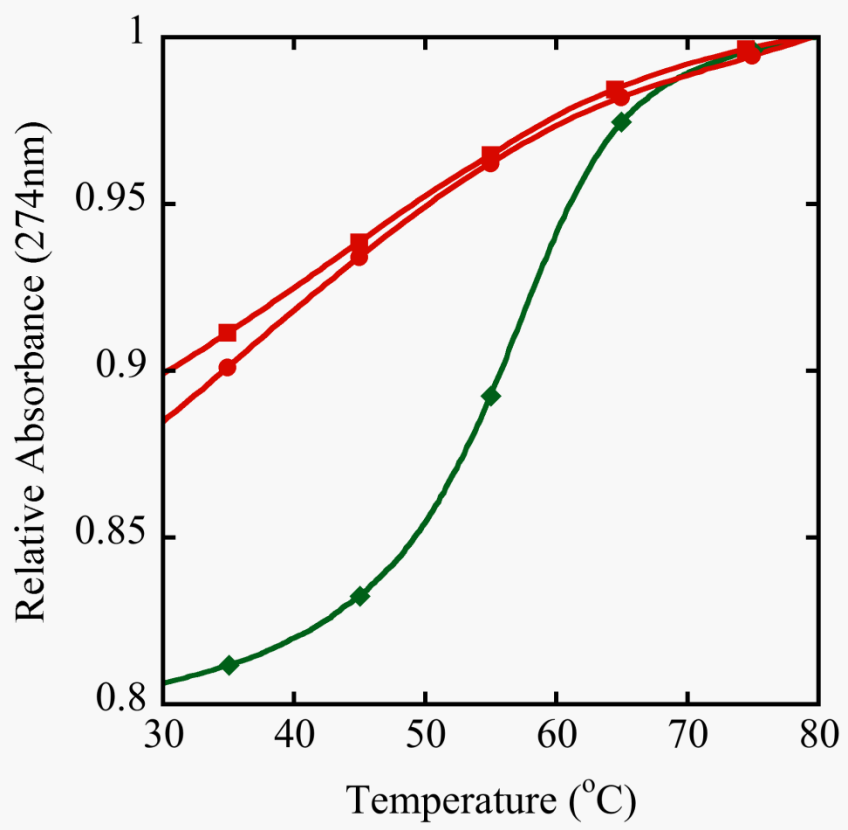


Figure 3.5 Duplex stability from UV Melting curves. Samples were prepared in 400 mM NaCl, 10 mM NaP_i, 0.1 mM EDTA at pH 6.5. The unmodified control decamer (II, ♦) at 8.5 µM showed a T_M of 59°C and two different concentrations of the selenium decamer (III) were compared, 8.5 µM (■) and 32 µM (●), both of which had an estimated T_M of 41°C.

hairpin structures. This data makes sense when compared to the imino proton spectra. The unmodified decamer **II** is shown to have five imino proton resonances, consistent with the C_2 rotational symmetry of a duplex structure (Figure 3.6A). In contrast sequence **III**, at 100 µM strand concentration, showed more imino proton resonances than would be expected for a duplex (Figure 3.6B). This strongly indicates the presence of multiple structures. Of note, there are resonances near 10.8 ppm that are generally associated with unpaired hairpin loop resonances.[42-43] If duplex **III** is examined at 20 µM strand concentration the spectrum simplifies and essentially only 3 GC base pairs are observed in addition to the hairpin loop resonances (Figure 3.6C). Under these

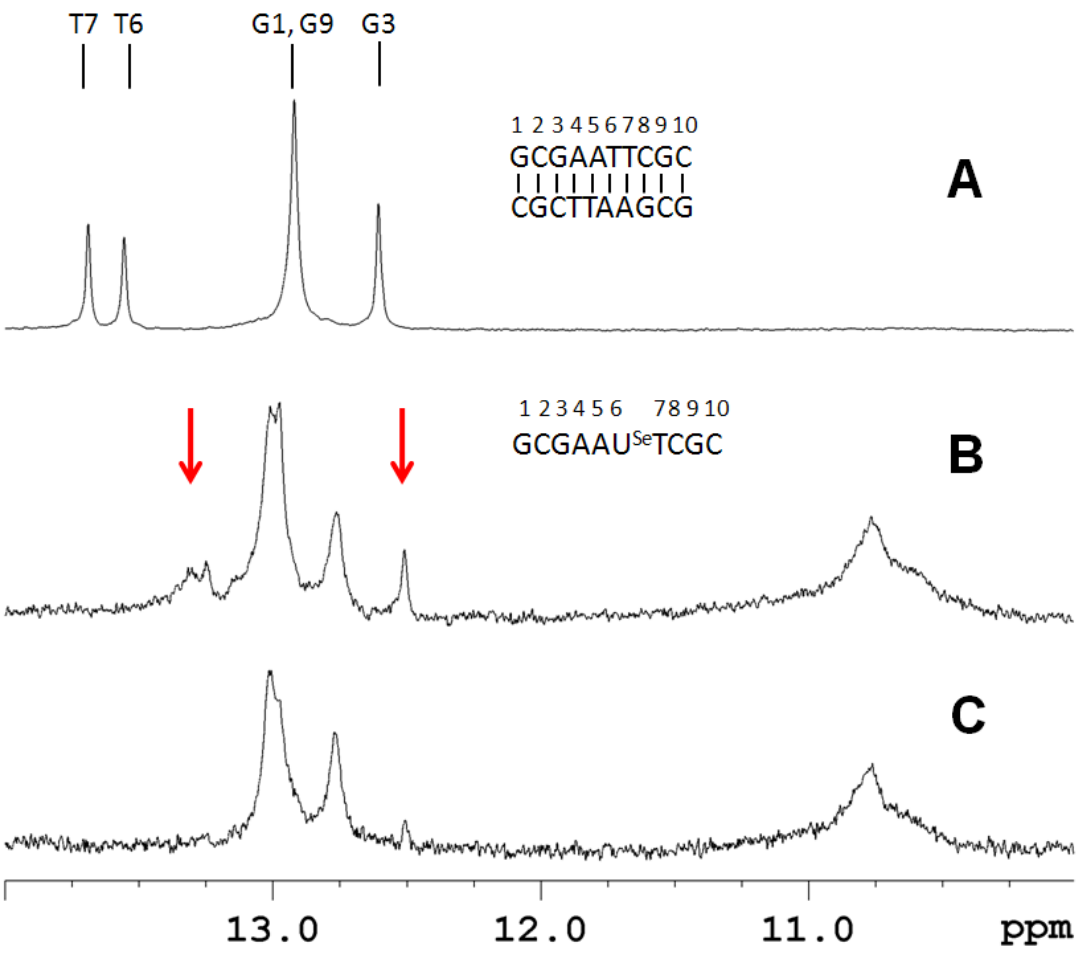


Figure 3.6 Imino proton spectra of sequences II and III at 288 K (A) The control decamer (**II**, 250 µM strand concentration) spectrum shows the presence of five base pairs. (B) and (C) are the selenium-containing decamer (**III**) at high and low strand concentration (100 and 20 µM, respectively). Arrows highlight resonances that disappear upon dilution. These signals are also sensitive to increased temperatures. Peaks are referenced to DSS.

conditions the predominant species is a hairpin structure with a stem consisting of 3 GC base pairs. Salon et al. have previously demonstrated complete base pairing for a non-self-complementary duplex containing a single 2'-Se-modification.[44] However the stability was compromised in this construct as well and homoduplex formation of the individual strands was observed at elevated temperatures. Taken in context, all of this data demonstrates that a 2'-selenomethyl group destabilizes a B-type DNA helix in

solution, even though this modification has an intrinsic preference for a southern sugar conformation, as demonstrated by the nucleoside study.

3.4 Crystal Structures and Models

Zhen Huang's group has made several selenium modified nucleosides and solved crystal structures with these analogs incorporated into a DNA helix (pdb: 1MA8, 3IFF, 2NSK, 2HC7, 2DLJ).[14,20] In every case, whether the 2'-selenomethyl-modification is present or not, the helix exhibits A-type, predominately N-sugar behavior. Even in 1MA8, where the selenomethyl groups are opposite and adjacent to each other, which was shown above to not exist in solution at room temperature, they situate themselves into the minor groove in a complete duplex (Figure 3.7a). Considering how deoxyuridine and 2'-methylseleno-uridine both prefer the S-conformation in solution, there has to be a driving force behind this change in overall sugar conformation. To obtain further insight why 2'-methylseleno-uridine (**6**) adopts a northern conformation when part of a DNA duplex, as seen in crystal structures, standard A- and B- type DNA helical models[45] with the appropriate modifications (i.e. sequence **III**) were investigated. The A-type helix containing the Se-modification was homologous to the crystal structures published in the literature. There are no steric clashes with the backbone, neighboring bases or deoxyribose ring. In contrast, in the B-helical model the modification is situated in the major groove, but the 3' phosphate as well as the base on the 3' side of the modification clash with selenomethyl group. This is especially apparent when the 3' base is thymine whose methyl group is also in the major groove. Therefore, a base with a smaller footprint in the major groove would be expected to be less perturbed which agrees with our previous NMR data where the 2' modified residue was flanked by cytosines.[44] To

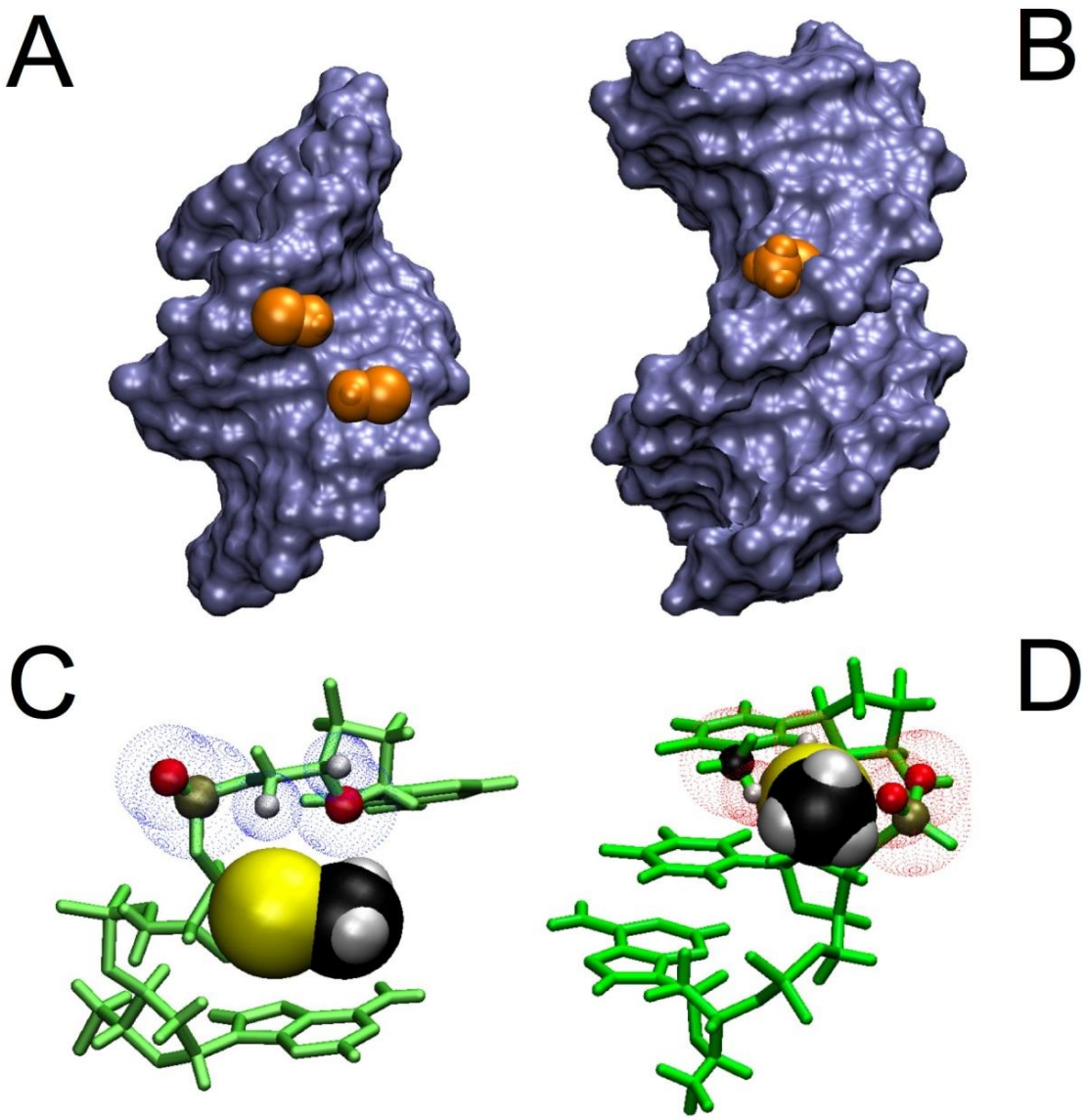


Figure 3.7 2' Se-CH_3 groups incorporated in A and B helical structures. Panels A and B represent Connelly surfaces with the Se-CH_3 group depicted as VdW representation in orange. In panel C and D the bases and sugars are shown in green and Se-CH_3 is depicted as VdW spheres. A) The group is nestled comfortably in the minor groove of an A-type helix (pdb: 1MA8). C) No clashes are apparent between the 3' neighboring residue and Se-CH_3 . The blue dotted spheres depict VdW spheres of close atoms. B) In a B-type helical model (dGCGAAU^{Se}TCGC) the Se-CH_3 group points away from the major groove but experiences significant clashes with the backbone as well as the base on the 3' side of the modification, which would disrupt base stacking. D) Predicted clashes for the B helix model are indicated with red VdW spheres for backbone (O and P) as well as the base (CH_3 and H6).

reiterate, the B-type model is intended to be a qualitative picture of how the selenium modification affects the stability of a B-type helix. There were no molecular dynamics simulations because there is no experimental data from which restraints could be obtained.

4 CONCLUSIONS

The notion that the southern sugar conformation is not tolerated well in a B-helix because of steric interactions explains the solution-based structural data but could also rationalize the enhanced crystal growth. DNA·RNA hybrids have been shown to form A-type structures, and alkylation of the 2'-hydroxyl group has been shown to increase the stability of these structures by lowering the intrinsic nucleophilicity of the hydroxyl group and altering the hydrogen bonding pattern. Addition of a methyl group no longer allows the hydroxyl group to act as a donor in a hydrogen bonding pair, and can now only accept hydrogen bonds. This effect has been said to drive local structure towards an A-type helix.[21] Divalent selenium atoms, as in compound **6**, can form hydrogen bonds with donor atoms, but because selenium has limited ability for induced dipoles due to its size, these bonds are very weak[46] and most likely has an influence on the hydrogen bonding network around the modification. Also, it has been proposed that the hydration of the minor groove will be affected by the pattern of purines and pyrimidines when a 2'-modification is made,[21] but recent data showed that the nature of the base of the modified nucleotide is not a determining factor as enhanced crystal growth was also observed for other 2'-selenomethyl-modified nucleotides.[47-49] Thus, the following can be concluded:

A single 2’-selenomethyl group narrows the conformational space by destabilizing the B-helical form while promoting A-helix formation. Moreover, the 2’-methylseleno group fits snuggly into the minor groove of an A-helix and can serve as the origin for a B- to A- conversion, which is also aided by dehydration during crystallization. In addition, the 2’-selenomethyl group locally dehydrates the minor groove which further facilitates the crystallization process. The impact of this behavior suggests that this modification could be used in samples that have been difficult to crystallize for their structural determination, yet consideration must be given to the fact that the conformational bias imparted by the modification could disrupt the normal behavior of the sample in solution. Positioning the modification in a place within a hairpin loop or on the 5’ side of a less-bulky purine base are the least likely to distort the structure by steric interactions.

5 REFERENCES

1 Altona, C. & Sundaralingam, M. Conformational analysis of the sugar ring in nucleosides and nucleotides. New description using the concept of pseudorotation. *JACS* **94**, 8205-8212, (1972).

2 Rich, A., Nordheim, A. & Wang, A. H. The chemistry and biology of left-handed Z-DNA. *Annu Rev Biochem* **53**, 791-846, (1984).

3 Sinden, R. R. *DNA Structure and Function*. (Academic Press, Inc., 1994).

4 Du, Y. H., Huang, J., Weng, X. C. & Zhou, X. Specific Recognition of DNA by Small Molecules. *Curr. Med. Chem.* **17**, 173-189 (2010).

5 Segal, D. J., Dreier, B., Beerli, R. R. & Barbas, C. F. Toward controlling gene expression at will: Selection and design of zinc finger domains recognizing each of the 5 '-GNN-3 ' DNA target sequences. *PNAS USA* **96**, 2758-2763 (1999).

6 Tolstorukov, M. Y., Jernigan, R. L. & Zhurkin, V. B. Protein-DNA hydrophobic recognition in the minor groove is facilitated by sugar switching. *J. Mol. Biol.* **337**, 65-76, (2004).

7 Hiratani, I. & Gilbert, D. M. Replication timing as an epigenetic mark. *Epigenetics* **4**, 93-97 (2009).

8 Kilpatrick, J. E., Pitzer, K. S. & Spitzer, R. The Thermodynamics and Molecular Structure of Cyclopentane. *JACS* **69**, 2483-2488, (1947).

9 Adams, W. J., Geise, H. J. & Bartell, L. S. Structure, equilibrium conformation, and pseudorotation in cyclopentane. An electron diffraction study. *JACS* **92**, 5013-5019, (1970).

10 Altona, C. & Sundaralingam, M. Conformational analysis of the sugar ring in nucleosides and nucleotides. Improved method for the interpretation of proton magnetic resonance coupling constants. *JACS* **95**, 2333-2344 (1973).

11 Karplus, M. Vicinal Proton Coupling in Nuclear Magnetic Resonance. *JACS* **85**, 2870-2871 (1963).

12 Vanwijk, J., Huckriede, B. D., Ippel, J. H. & Altona, C. Furanose Sugar Conformations in DNA from NMR Coupling Constants. *Methods Enzymol.* **211**, 286-306 (1992).

13 Leslie, A. G., Arnott, S., Chandrasekaran, R. & Ratliff, R. L. Polymorphism of DNA double helices. *J Mol Biol* **143**, 49-72, (1980).

14 Jiang, J., Sheng, J., Carrasco, N. & Huang, Z. Selenium derivatization of nucleic acids for crystallography. *Nucleic Acids Res* **35**, 477-485 (2007).

15 Meisenheimer, K. M. & Koch, T. H. Photocross-linking of nucleic acids to associated proteins. *Crit. Rev. Biochem. Mol. Biol.* **32**, 101-140 (1997).

16 Hendrickson, W. A., Horton, J. R. & LeMaster, D. M. Selenomethionyl proteins produced for analysis by multiwavelength anomalous diffraction (MAD): a vehicle for direct determination of three-dimensional structure. *EMBO J* **9**, 1665-1672 (1990).

17 Du, Q. *et al.* Internal Derivatization of Oligonucleotides with Selenium for X-ray Crystallography Using MAD. *JACS* **124**, 24-25, (2001).

18 Sheng, J., Jiang, J., Salon, J. & Huang, Z. Synthesis of a 2'-Se-thymidine phosphoramidite and its incorporation into oligonucleotides for crystal structure study. *Org Lett* **9**, 749-752, (2007).

19 Sheng, J. & Huang, Z. Selenium derivatization of nucleic acids for X-ray crystal-structure and function studies. *Chem Biodivers* **7**, 753-785, (2010).

20 Sheng, J. & Huang, Z. Selenium Derivatization of Nucleic Acids for Phase and Structure Determination in Nucleic Acid X-ray Crystallography. *Int J Mol Sci* **9**, 258-271 (2008).

21 Lubini, P., Zürcher, W. & Egli, M. Stabilizing effects of the RNA 2'-substituent: crystal structure of an oligodeoxynucleotide duplex containing 2'-O-methylated adenosines. *Chem. Biol.* **1**, 39-45 (1994).

22 Rozners, E. Carbohydrate Chemistry for RNA Interference: Synthesis and Properties of RNA Analogues Modified in Sugar-Phosphate Backbone. *Curr. Org. Chem.* **10**, 675-692 (2006).
23 Fraser, A., Wheeler, P., Cook, P. D. & Sanghvi, Y. S. Synthesis and conformational properties of 2'-deoxy-2'-methylthiopyrimidine and 2'-deoxy-2'methylthio-purine nucleosides: Potential antisense applications. *J. Heterocycl. Chem.* **30**, 1277-1287 (1993).
24 de Leeuw, F. A. A. M., Altona, C. Computer-Assisted Pseudorotation Analysis of Five-membered Rings by Means of Proton Spin-Spin Coupling Constants: Program PSEUROT. *J. Comput. Chem.* **4**, 428-437 (1983).
25 Hendrickx, P. & Martins, J. A user-friendly Matlab program and GUI for the pseudorotation analysis of saturated five-membered ring systems based on scalar coupling constants. *Chemistry Central Journal* **2**, 20 (2008).
26 Houseknecht, J. B., Altona, *et al.* Conformational Analysis of Furanose Rings with PSEUROT: Parametrization for Rings Possessing the Arabino, Lyxo, Ribo, and Xylo Stereochemistry and Application to Arabinofuranosides. *J. Org. Chem.* **67**, 4647-4651, (2002).
27 Rosemeyer, H. *et al.* Stereoelectronic effects of modified purines on the sugar conformation of nucleosides and fluorescence properties. *Nucleosides Nucleotides Nucl. Acids* **16**, 821-828 (1997).
28 Watts, J. K., Sadalapure, K., *et al.* J. Synthesis and Conformational Analysis of 2‘-Fluoro-5-methyl-4‘-thioarabinouridine (4‘S-FMAU). *J. Org. Chem.* **71**, 921-925 (2006).
29 Altona, C. *et al.* Empirical group electronegativities for vicinal NMR proton-proton couplings along a C-C bond: solvent effects and reparameterization of the Haasnoot equation. *Magn. Reson. Chem.* **32**, 670-678 (1994).
30 Altona, C. *et al.* Relationship between proton-proton NMR coupling constants and substituent electronegativities. V. Empirical substituent constants deduced from ethanes and propanes. *Magn. Reson. Chem.* **27**, 564-576 (1989).
31 Haasnoot, C. A. G., De, L. F. A. A. M., De, L. H. P. M. & Altona, C. The relationship between proton-proton NMR coupling constants and substituent electronegativities. Part II. Conformational analysis of the sugar ring in nucleosides and nucleotides in solution using a generalized Karplus equation. *Org. Magn. Reson.* **15**, 43-52 (1981).
32 Donders, L. A., Leeuw, F. A. A. M. d. & Altona, C. Relationship Between Proton-Proton NMR Coupling Constants and Subsituent Electronegativites IV: An Extended Karplus Equation Accounting for Interactions Between Substituents and its Application to Coupling Constant Data Calculated by the Extended Hueckel Method. *Magn. Reson. Chem.* **27**, 556-563 (1989).
33 Plateau, P. & Gueron, M. Exchangeable proton NMR without base-line distorsion, using new strong-pulse sequences. *JACS* **104**, 7310-7311 (1982).
34 Davies, D. B. & Danyluk, S. S. Nuclear magnetic resonance studies of 5'-ribo- and deoxyribonucleotide structures in solution. *Biochemistry* **13**, 4417-4434 (1974).
35 Guschlbauer, W. & Jankowski, K. Nucleoside Conformation is determined by the Electronegativity of the Sugar Substituent. *Nucleic Acids Res* **8**, 1421-1433 (1980).
36 Joecks, A., Koeppel, H., Schleinitz, K. D. & Cech, D. NMR-spektroskopische Untersuchengen zum Konformationsverhalten von 2'- und 3'-halogensubstituierten Pyrimidinnucleosiden. *Journal fuer praktische Chemie* **325**, 881 (1983).
37 Robins, M. J., *et al.* Nucleic acid related compounds. 73. Fluorination of uridine 2'-thioethers with xenon difluoride or (diethylamino)sulfur trifluoride. Synthesis of stable 2'-[alkyl(or aryl)sulfonyl]-2'-deoxy-2'-fluorouridines. *J. Org. Chem.* **57**, 2357-2364 (1992).

38 Uesugi, S., Miki, H., Ikehara, M., Iwahashi, H. & Kyogoku, Y. A linear relationship between electronegativity of 2'-substituents and conformation of adenine nucleosides. *Tetrahedron Lett.* **20**, 4073-4076 (1979).

39 Rinkel, L. T. & Altona, C. Conformational Analysis of the Deoxyribofuranose Ring in DNA by means of sums of proton-proton coupling constants: A Graphical Method. *J. Biomol. Struct. Dyn.* **4**, 621-649 (1987).

40 Aramini, J. M., Kalisch, B. W., Pon, R. T., vandeSande, J. H. & Germann, M. W. Structure of a DNA duplex that contains alpha-anomeric nucleotides and 3'-3' and 5'-5' phosphodiester linkages: Coexistence of parallel and antiparallel DNA. *Biochemistry* **35**, 9355-9365 (1996).

41 Aramini, J. M., Mujeeb, A. & Germann, M. W. NMR solution structures of d(GCGAAT-3 '-3 '-alpha T-5 '-5 '-CGC)(2) and its unmodified control. *Nucleic Acids Res.* **26**, 5644-5654 (1998).

42 Germann, M. W., Kalisch, B. W., Lundberg, P., Vogel, H. J. & van de Sande, J. H. Perturbation of DNA hairpins containing the EcoRI recognition site by hairpin loops of varying size and composition: physical (NMR and UV) and enzymatic (EcoRI) studies. *Nucleic Acids Res.* **18**, 1489-1498 (1990).

43 Haasnoot, C. A. G., der Hartog, J. H. J., de Rooij, J. F. M., van Boom, J. H. & Altona, C. Loopstructures in synthetic oligonucleotides. *Nucleic Acids Res.* **8**, 169-181 (1980).

44 Salon, J., Chen, G., Portilla, Y., Germann, M. W. & Huang, Z. Synthesis of a 2‘-Se-uridine Phosphoramidite and Its Incorporation into Oligonucleotides for Structural Study. *Org. Lett.* **7**, 5645-5648 (2005).

45 Egli, M. Structural Aspects of Nucleic Acid Analogs and Antisense Oligonucleotides. *Angewandte Chemie International Edition in English* **35**, 1894-1909 (1996).

46 Madzhidov, T. I. & Chmutova, G. A. The nature of hydrogen bonds with divalent selenium compounds. *Journal of Molecular Structure: THEOCHEM* **959**, 1-7 (2010).

47 Salon, J., Sheng, J., Gan, J. & Huang, Z. Synthesis and crystal structure of 2'-Se-modified guanosine containing DNA. *J Org Chem* **75**, 637-641 (2010).

48 Sheng, J., Salon, J., Gan, J. H. & Huang, Z. Synthesis and crystal structure study of 2 '-Se-adenosine-derivatized DNA. *Science China-Chemistry* **53**, 78-85 (2010).

49 Buzin, Y., Carrasco, N. & Huang, Z. Synthesis of Selenium-Derivatized Cytidine and Oligonucleotides for X-ray Crystallography Using MAD. *Org. Lett.* **6**, 1099-1102 (2004).

6 Appendix

6.1 Appendix A - Reconstructed PSEUROT Batch File

```
::This batch file was reconstructed from the version contained in a degraded copy of PSEUROT
::that had been copied multiple times over a few years. As I am not a computer programmer,
::I'm not entirely sure why it works, but it does. When using the command line context
::described in the manual, this .bat file correctly renames inputs and outputs for use in the
::PSEUROT program, and the output files are competent. PSEUROT is able to run to completion,
::which is stated at the end of the output files. However, the 'MANY' functionality does not
::work to completion.

@ECHO OFF
if '%1'=='' goto Usage
copy %1 %1.inp
copy %1.inp pseurot6.inp
psrot62 <pseurot6.inp >pseurot6.out
if exist pseurot6.out    copy pseurot6.out %1.out >NUL
if exist pseurot6.mn1    copy pseurot6.mn1 %1.mn1 >NUL
if exist pseurot6.mn2    copy pseurot6.mn2 %1.mn2 >NUL
if exist pseurot6.mn3    copy pseurot6.mn3 %1.mn3 >NUL
if exist pseurot6.mn4    copy pseurot6.mn4 %1.mn4 >NUL
if exist pseurot6.mn5    copy pseurot6.mn5 %1.mn5 >NUL
:pkzip %1.zip %1.inp %1.mn1 %1.mn2 %1.mn3 %1.mn4 %1.mn5 :goto Einde
:Usage
       echo Usage: PS62 filename
       echo where filename does not have an extension
       echo.
:Einde
```

6.2 Appendix B – Inputs and Results for PSEUROT 6.0 Calculations

6.2.1 2'-Deoxyuridine

Trial 1 Input

```
dU
CTRL  MAXIT 25  TRIM 0.1  RCNV 0.5  PRINT 1
DATA 5
1'-2'    -144.0       1.030   121.4       0.72      1.27      0.00      0.62
1'-2"    -144.0       1.020     0.9       0.72      1.27      0.62      0.00
2'-3'       0.0       1.060     2.4       0.62      0.00      1.26      0.62
2"-3'       0.0       1.060   122.9       0.00      0.62      1.26      0.62
3'-4'     144.0       1.090  -124.0       0.72      1.26      1.27      0.68
TSET 1
298       7.19  6.14  3.89  6.9  3.89
START  26.0      38.0     164.0      38.0   .78
FITF 00111
```

Trial 1 Output

```
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
+  PSEUROT v 6.0                          March 1993  +
+  John van Wijk                        FAAM de Leeuw  +
+  Gorlaeus Laboratories, State University of Leiden  +
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

++++ CASE :  1 ++++++++++++++++++++++++++++++++++++++++++++ PSEUROT 6.0 +++++

TITLE:
dU
The minimization has converged.
================================================================================
 F I N A L   O U T P U T
================================================================================
Total number of iterations:  9

CONFORMER I:                              CONFORMER II:

      P ==    18.0  (   .314 RAD)             P ==   184.6  ( 3.222 RAD)
   PHIM ==    38.0  (   .663 RAD)          PHIM ==    59.7  ( 1.041 RAD)

  PHIHH =    98.4    ==> JHH =   1.22    PHIHH =   168.0    ==> JHH =  11.24
  PHIHH =   -21.9    ==> JHH =   7.70    PHIHH =    47.1    ==> JHH =   4.11
  PHIHH =    40.7    ==> JHH =   6.56    PHIHH =   -60.6    ==> JHH =   2.04
  PHIHH =   161.2    ==> JHH =  10.01    PHIHH =    59.9    ==> JHH =   3.50
  PHIHH =  -163.4    ==> JHH =   8.81    PHIHH =   -68.5    ==> JHH =   1.58

TEMP SET                  298
            JEXP   JCAL   JDIF
   1'-2'    7.19   7.07    .12
   1'-2"    6.14   5.60    .54
   2'-3'    3.89   3.92   -.03
   2"-3'    6.90   6.21    .69
   3'-4'    3.89   4.59   -.70
X(1)X(2)     .42    .58   .505

ERROR ANALYSIS:
ROOT MEAN SQUARE DEVIATION OF THE FIT:      .505
STANDARD DEVIATION IN PARAMETERS:
    0      .208      .112      .061

CORRELATION MATRIX OF PARAMETERS

PAR.       1          2          3
 1      1.000
 2       .305      1.000
 3       .509       .212      1.000

END OF THE PROGRAM PSEUROT 6.0
```

Trial 2 Input

```
dU_databse parameters
CTRL  MAXIT 25  TRIM 0.1  RCNV 0.5  PRINT 1
DATA 5
1'-2'    -144.0       1.030   121.4       0.56      1.26      0.00      0.62
1'-2"    -144.0       1.020     0.9       0.56      1.26      0.62      0.00
2'-3'       0.0       1.060     2.4       0.62      0.00      1.26      0.62
2"-3'       0.0       1.060   122.9       0.00      0.62      1.26      0.62
3'-4'     144.0       1.090  -124.0       0.67      1.26      1.26      0.68
TSET 1
298       7.19  6.14  3.89  6.9  3.89
START  18.0      38.0     162.0      33.0   .78
FITF 00111
```

Trial 2 Output

```
The minimization has converged.
================================================================================
 F I N A L   O U T P U T
================================================================================
Total number of iterations:  8

CONFORMER I:                              CONFORMER II:

      P ==    18.0  (   .314 RAD)             P ==   181.0  ( 3.159 RAD)
   PHIM ==    38.0  (   .663 RAD)          PHIM ==    59.1  ( 1.032 RAD)

  PHIHH =    98.4    ==> JHH =   1.29    PHIHH =   170.0    ==> JHH =  11.63
  PHIHH =   -21.9    ==> JHH =   8.03    PHIHH =    49.0    ==> JHH =   3.87
  PHIHH =    40.7    ==> JHH =   6.56    PHIHH =   -60.2    ==> JHH =   2.08
  PHIHH =   161.2    ==> JHH =  10.01    PHIHH =    60.3    ==> JHH =   3.45
  PHIHH =  -163.4    ==> JHH =   8.86    PHIHH =   -71.2    ==> JHH =   1.43

TEMP SET                  298
            JEXP   JCAL   JDIF
   1'-2'    7.19   7.18    .01
   1'-2"    6.14   5.66    .48
   2'-3'    3.89   4.01   -.12
   2"-3'    6.90   6.27    .63
   3'-4'    3.89   4.63   -.74
X(1)X(2)     .43    .57   .487

ERROR ANALYSIS:
ROOT MEAN SQUARE DEVIATION OF THE FIT:      .487
STANDARD DEVIATION IN PARAMETERS:
    0      .224      .107      .057

CORRELATION MATRIX OF PARAMETERS

PAR.       1          2          3
 1      1.000
 2       .187      1.000
 3       .503       .148      1.000

END OF THE PROGRAM PSEUROT 6.0
```

6.2.2 Uridine

Trial 1 Input

```
rU_database_parameters
ctrl  maxit 25  trim 0.1  rcnv 0.5  print 1
data 3
1'-2'      -144.0       1.102    123.3       0.56      1.26      1.26      0.62
2'-3'         0.0       1.090      0.2       0.62      1.26      1.26      0.62
3'-4'       144.0       1.095   -124.9       0.62      1.26      1.26      0.68
tset 1
298       4.5  5.3  5.5
start  18.0      32.0     153.6      35.0   .20
fitf 10101
```

Trial 1 Output

```
rU_database_parameters
MAXIMUM NUMBER OF ITERATIONS 25 REACHED.

================================================================================
 F I N A L   O U T P U T
================================================================================
Total number of iterations: 25

CONFORMER I:                            CONFORMER II:

      P ==    32.4  (   .566 RAD)            P ==   156.6  ( 2.733 RAD)
   PHIM ==    32.0  (   .559 RAD)         PHIM ==    35.0  (   .611 RAD)

  PHIHH =   110.3    ==> JHH =   1.53    PHIHH =   160.9    ==> JHH =   8.89
  PHIHH =    29.6    ==> JHH =   5.47    PHIHH =   -34.8    ==> JHH =   5.03
  PHIHH =  -159.9    ==> JHH =   8.53    PHIHH =  -105.4    ==> JHH =   1.04

TEMP SET                  298
           JEXP   JCAL   JDIF
   1'-2'   4.50   4.50    .00
   2'-3'   5.30   5.29    .01
   3'-4'   5.50   5.50    .00
X(1)X(2)    .60    .40   .004

ROOT MEAN SQUARE DEVIATION OF THE FIT:      .004


FIT   3 OBS TO   3 PARS -> ERROR ANALYSIS OMITTED

END OF THE PROGRAM PSEUROT 6.0
```

6.2.3 2'-Methoxy-Uridine

Trial 1 Input

```
OMe_database_parameters
ctrl  maxit 25  trim 0.1  rcnv 0.5  print 1
data 3
1'-2'      -144.0       1.102    123.3       0.56      1.26      1.26      0.62
2'-3'         0.0       1.090      0.2       0.62      1.26      1.26      0.62
3'-4'       144.0       1.095   -124.9       0.62      1.26      1.26      0.68
tset 1
298       3.9  5.2  5.7
start  18.0      32.0     153.6      35.0   .20
fitf 10101
```

Trial 1 Output

```
OMe_database_parameters
MAXIMUM NUMBER OF ITERATIONS 25 REACHED.

================================================================================
 F I N A L   O U T P U T
================================================================================
Total number of iterations: 25

CONFORMER I:                            CONFORMER II:

      P ==    12.5  (   .218 RAD)            P ==   144.7  ( 2.526 RAD)
   PHIM ==    32.0  (   .559 RAD)         PHIM ==    35.0  (   .611 RAD)

  PHIHH =    99.9    ==> JHH =    .80    PHIHH =   161.9    ==> JHH =   8.99
  PHIHH =    34.3    ==> JHH =   5.08    PHIHH =   -30.9    ==> JHH =   5.37
  PHIHH =  -157.0    ==> JHH =   8.18    PHIHH =  -112.6    ==> JHH =   1.64

TEMP SET                  298
           JEXP   JCAL   JDIF
   1'-2'   3.90   3.90    .00
   2'-3'   5.20   5.19    .01
   3'-4'   5.70   5.70    .00
X(1)X(2)    .62    .38   .007

ROOT MEAN SQUARE DEVIATION OF THE FIT:      .007


FIT   3 OBS TO   3 PARS -> ERROR ANALYSIS OMITTED

END OF THE PROGRAM PSEUROT 6.0
```

6.2.4 2'-Fluoro-Deoxyuridine

Trial 1 Input

```
::F, trial 1, electronegativity changes based on table V B in full description

2F_database_parameters
ctrl  maxit 25  trim 0.1  rcnv 0.5  print 1
data 3
1'-2'      -144.0       1.102   123.3       0.56      1.26      1.37      0.62
2'-3'        0.0        1.090    0.2        0.62      1.37      1.26      0.62
3'-4'      144.0        1.095   -124.9      0.62      1.26      1.26      0.68
tset 1
298       1.38  5.0  8.6
start  18.0      32.0     153.6      35.0   .20
fitf 10101
```

Trial 1 Output

```
2F_database_parameters
The minimization has converged.

================================================================================
 F I N A L   O U T P U T
================================================================================
Total number of iterations:  9

CONFORMER I:                            CONFORMER II:

      P ==    30.5  (   .533 RAD)            P ==   150.6  ( 2.629 RAD)
   PHIM ==    32.0  (   .559 RAD)         PHIM ==    35.0  (   .611 RAD)

  PHIHH =   109.3    ==> JHH =   1.28    PHIHH =   161.6    ==> JHH =   8.71
  PHIHH =    30.2    ==> JHH =   5.14    PHIHH =   -33.0    ==> JHH =   5.26
  PHIHH =  -159.8    ==> JHH =   8.52    PHIHH =  -108.9    ==> JHH =   1.30

TEMP SET                   298
           JEXP   JCAL   JDIF
   1'-2'   1.38   1.28    .10
   2'-3'   5.00   5.14   -.14
   3'-4'   8.60   8.52    .08
X(1)X(2)   1.00    .00   .111

ROOT MEAN SQUARE DEVIATION OF THE FIT:      .111


FIT   3 OBS TO   3 PARS -> ERROR ANALYSIS OMITTED

END OF THE PROGRAM PSEUROT 6.0
```

Trial 2 Input

```
::"F, trial 2, predominately S"

F_database_parameters_predom_S
ctrl  maxit 25  trim 0.1  rcnv 0.5  print 1
data 3
1'-2'      -144.0       1.102   123.3       0.56      1.26      1.37      0.62
2'-3'         0.0       1.090    0.2       0.62      1.37      1.26      0.62
3'-4'       144.0       1.095   -124.9       0.62      1.26      1.26      0.68
tset 1
298       1.4  5.0  8.6
start  18.0      32.0     153.6      35.0   .80
fitf 10101
```

Trial 2 Output

```
F_database_parameters_predom_S
MAXIMUM NUMBER OF ITERATIONS 25 REACHED.

================================================================================
 F I N A L   O U T P U T
================================================================================
Total number of iterations: 25

CONFORMER I:                              CONFORMER II:

      P ==    28.6  (   .499 RAD)             P ==    38.6  (   .673 RAD)
   PHIM ==    32.0  (   .559 RAD)          PHIM ==    35.0  (   .611 RAD)

  PHIHH =   108.2    ==> JHH =   1.19    PHIHH =   113.0    ==> JHH =   1.63
  PHIHH =    30.8    ==> JHH =   5.09    PHIHH =    30.0    ==> JHH =   5.16
  PHIHH =  -159.6    ==> JHH =   8.50    PHIHH =  -163.2    ==> JHH =   8.89

TEMP SET                  298
           JEXP   JCAL   JDIF
   1'-2'   1.40   1.33    .07
   2'-3'   5.00   5.11   -.11
   3'-4'   8.60   8.63   -.03
X(1)X(2)    .68    .32   .079

ROOT MEAN SQUARE DEVIATION OF THE FIT:      .079


FIT   3 OBS TO   3 PARS -> ERROR ANALYSIS OMITTED

END OF THE PROGRAM PSEUROT 6.0
```

6.2.5 2'-Methylthio-Deoxyuridine

Trial 1 Input

```
::"SMe, trial 1, electronegativity from table V B"

SMe_database_parameters
ctrl  maxit 25  trim 0.1  rcnv 0.5  print 1
data 3
1'-2'      -144.0       1.102    123.3       0.56      1.26      0.7      0.62
2'-3'         0.0       1.090      0.2       0.62      0.7      1.26      0.62
3'-4'       144.0       1.095   -124.9       0.62      1.26      1.26      0.68
tset 1
298       8.3  5.7  2.83
start  18.0      32.0     153.6      35.0   .20
fitf 10101
```

Trial 1 Output

```
SMe_database_parameters
MAXIMUM NUMBER OF ITERATIONS 25 REACHED.

================================================================================
 F I N A L   O U T P U T
================================================================================
Total number of iterations: 25

CONFORMER I:                            CONFORMER II:

      P ==   -22.5  ( -.393 RAD)            P ==   138.7  ( 2.421 RAD)
   PHIM ==    32.0  (  .559 RAD)         PHIM ==    35.0  (  .611 RAD)

  PHIHH =    89.0    ==> JHH =    .69    PHIHH =   161.7    ==> JHH =  10.08
  PHIHH =    32.4    ==> JHH =   6.46    PHIHH =   -28.5    ==> JHH =   5.52
  PHIHH =  -143.2    ==> JHH =   6.12    PHIHH =  -116.5    ==> JHH =   2.06

TEMP SET                  298
            JEXP   JCAL   JDIF
   1'-2'    8.30   8.30    .00
   2'-3'    5.70   5.70    .00
   3'-4'    2.83   2.83    .00
X(1)X(2)     .19    .81   .000

ROOT MEAN SQUARE DEVIATION OF THE FIT:      .000


FIT   3 OBS TO   3 PARS -> ERROR ANALYSIS OMITTED

END OF THE PROGRAM PSEUROT 6.0
```

Trial 2 Input

```
::"SMe, trial 2, predominately S starting cond"

SMe_database_parameters_predom_S
ctrl  maxit 25  trim 0.1  rcnv 0.5  print 1
data 3
1'-2'      -144.0       1.102   123.3       0.56      1.26      0.7      0.62
2'-3'         0.0       1.090    0.2       0.62      0.7      1.26      0.62
3'-4'       144.0       1.095   -124.9       0.62      1.26      1.26      0.68
tset 1
298       8.3  5.8  2.8
start  18.0      32.0     153.6      35.0   .80
fitf 10101
```

Trial 2 Output

```
SMe_database_parameters_predom_S
MAXIMUM NUMBER OF ITERATIONS 25 REACHED.

================================================================================
 F I N A L   O U T P U T
================================================================================
Total number of iterations: 25

CONFORMER I:                              CONFORMER II:

      P ==   -30.8  ( -.538 RAD)             P ==   136.9  ( 2.390 RAD)
   PHIM ==    32.0  (  .559 RAD)          PHIM ==    35.0  (  .611 RAD)

  PHIHH =    88.2    ==> JHH =    .67    PHIHH =   161.6    ==> JHH =  10.07
  PHIHH =    30.2    ==> JHH =   6.65    PHIHH =   -27.7    ==> JHH =   5.60
  PHIHH =  -138.7    ==> JHH =   5.38    PHIHH =  -117.6    ==> JHH =   2.20

TEMP SET                   298
           JEXP   JCAL   JDIF
   1'-2'   8.30   8.30    .00
   2'-3'   5.80   5.80    .00
   3'-4'   2.80   2.80    .00
X(1)X(2)    .19    .81   .000

ROOT MEAN SQUARE DEVIATION OF THE FIT:      .000


FIT   3 OBS TO   3 PARS -> ERROR ANALYSIS OMITTED

END OF THE PROGRAM PSEUROT 6.0
```

6.2.6 2'-Selenomethyl-Deoxyuridine

Trial 1 Input

```
::"SeMe, trial 1, the electronegativity for S from table V B used, since it was an extra-
::polated value and the Pauling negativity of S and Se only differ by 0.03. The Pauling
::scale is broader than the Altona scale, which is used here and is correlated to coup-
::ling constants

SeMe_database_parameters
ctrl  maxit 25  trim 0.1  rcnv 0.5  print 1
data 3
1'-2'      -144.0       1.102   123.3       0.56      1.26      0.68      0.62
2'-3'         0.0       1.090    0.2       0.62      0.68      1.26      0.62
3'-4'       144.0       1.095   -124.9       0.62      1.26      1.26      0.68
tset 1
298       8.65  5.7  2.86
start  18.0      32.0     153.6      35.0   .20
fitf 10101
```

Trial 1 Output

```
SeMe_database_parameters
MAXIMUM NUMBER OF ITERATIONS 25 REACHED.

================================================================================
 F I N A L   O U T P U T
================================================================================
Total number of iterations: 25

CONFORMER I:                            CONFORMER II:

      P ==   -14.8  ( -.258 RAD)             P ==   137.6  ( 2.402 RAD)
   PHIM ==    32.0  (  .559 RAD)          PHIM ==    35.0  (  .611 RAD)

  PHIHH =    90.4    ==> JHH =    .72    PHIHH =   161.6    ==> JHH =  10.11
  PHIHH =    33.9    ==> JHH =   6.37    PHIHH =   -28.0    ==> JHH =   5.58
  PHIHH =  -147.1    ==> JHH =   6.74    PHIHH =  -117.2    ==> JHH =   2.15

TEMP SET                  298
            JEXP   JCAL   JDIF
   1'-2'    8.65   8.65    .00
   2'-3'    5.70   5.70    .00
   3'-4'    2.86   2.86    .00
X(1)X(2)     .16    .84   .000

ROOT MEAN SQUARE DEVIATION OF THE FIT:      .000


FIT   3 OBS TO   3 PARS -> ERROR ANALYSIS OMITTED

END OF THE PROGRAM PSEUROT 6.0
```

Trial 2 Input

```
::"SeMe, trial 2, starting predom S"

SeMe_database_parameters_predom_S
ctrl  maxit 25  trim 0.1  rcnv 0.5  print 1
data 3
1'-2'      -144.0       1.102   123.3       0.56      1.26      0.68      0.62
2'-3'         0.0       1.090    0.2       0.62      0.68      1.26      0.62
3'-4'       144.0       1.095   -124.9       0.62      1.26      1.26      0.68
tset 1
298       8.6  5.7  2.9
start  18.0      32.0     153.6      35.0   .80
fitf 10101
```

Trial 2 Output

```
SeMe_database_parameters_predom_S
MAXIMUM NUMBER OF ITERATIONS 25 REACHED.

================================================================================
 F I N A L   O U T P U T
================================================================================
Total number of iterations: 25

CONFORMER I:                            CONFORMER II:

      P ==   -13.3  ( -.233 RAD)             P ==   137.6  ( 2.402 RAD)
   PHIM ==    32.0  (  .559 RAD)          PHIM ==    35.0  (  .611 RAD)

  PHIHH =    90.8    ==> JHH =    .73    PHIHH =   161.6    ==> JHH =  10.11
  PHIHH =    34.1    ==> JHH =   6.35    PHIHH =   -28.0    ==> JHH =   5.58
  PHIHH =  -147.7    ==> JHH =   6.85    PHIHH =  -117.2    ==> JHH =   2.15

TEMP SET                  298
           JEXP   JCAL   JDIF
   1'-2'   8.60   8.60    .00
   2'-3'   5.70   5.70    .00
   3'-4'   2.90   2.90    .00
X(1)X(2)    .16    .84   .000

ROOT MEAN SQUARE DEVIATION OF THE FIT:      .000


FIT   3 OBS TO   3 PARS -> ERROR ANALYSIS OMITTED

END OF THE PROGRAM PSEUROT 6.0
```

6.3 Appendix C - Compiled Matlab Results

6.3.1 2'-Deoxyuridine

```
////////////////////////////////////////////////////////////

  START Pseudorotational calculation

////////////////////////////////////////////////////////////
Local minimum possible. Constraints satisfied.

No active inequalities.

------------------------------------------------------------
  Optimized parameters
------------------------------------------------------------

   Conformation 1
    P     :    2.314
    Phi_m :   33.513

   Conformation 2
    P     :  149.771
    Phi_m :   22.999

   Temperature Coefficients
    %Conformation1  :   34.682


------------------------------------------------------------
  Endocyclic torsion angles
------------------------------------------------------------
```

	Conf.1	Conf.2
Phi0:	33.500	-19.602
Phi1:	-27.455	8.793
Phi2:	10.924	5.374
Phi3:	9.781	-17.489
Phi4:	-26.749	22.924

```
------------------------------------------------------------
  Final couplings
------------------------------------------------------------

   Temperature 1:
    ------------------------------------
    Conf1  Conf2    Avg.   Exp.  Diff.
    ------------------------------------
     7.70   6.92   7.19   7.19   -0.00
     0.99   8.88   6.14   6.14    0.00
     6.84   6.93   6.90   6.90    0.00
     9.64   0.84   3.89   3.89   -0.00
     7.71   1.87   3.89   3.89   -0.00
    ------------------------------------
    RMSD :     0.00 Hz


------------------------------------------------------------
  ERROR ANALYSIS
------------------------------------------------------------

  i       dP1/dPi     dP2/dPi     dP3/dPi     dP4/dPi     dP5/dPi    dRMSD/dPi
  1         1.000      -0.704       0.566       0.380       0.405        0.015
  2        -0.625       1.000      -0.415      -0.470      -0.485        0.015
  3         1.267      -0.955       1.000       0.531       0.574        0.021
  4         1.294      -1.724       0.745       1.000       1.038        0.025
  5         1.272      -1.677       0.745       0.949       1.000        0.023

  i       dP1/dCi     dP2/dCi     dP3/dCi     dP4/dCi     dP5/dCi    dRMSD/dCi
  1        80.440     -34.693      73.888      35.526      56.410        0.017
  2        95.746     -37.006      34.880      46.578      64.161        0.024
  3        -9.106     -27.014     -23.388      14.642      27.562        0.058
  4        62.868     -38.223      32.906      51.052      66.891        0.067
  5       119.476     -38.679      25.426      74.002      87.071        0.153

------------------------------------------------------------
  RMSD                :       0.00 Hz
  Total time          :       0.92 s
------------------------------------------------------------
```

6.3.2 Uridine

```
////////////////////////////////////////////////////////////

  START Pseudorotational calculation

////////////////////////////////////////////////////////////
////////////////////////////////////////////////////////////


Local minimum possible. Constraints satisfied.

No active inequalities.


------------------------------------------------------------
  Optimized parameters
------------------------------------------------------------

   Conformation 1
    P     :  169.600
    Phi_m :   30.892

   Conformation 2
    P     :   46.669
    Phi_m :   39.158

   Temperature Coefficients
    %Conformation1  :   43.120

------------------------------------------------------------
  Endocyclic torsion angles
------------------------------------------------------------
```

	Conf.1	Conf.2
Phi0:	-30.377	27.327
Phi1:	21.043	-38.670
Phi2:	-3.671	35.242
Phi3:	-15.103	-18.353
Phi4:	28.108	-5.546

```
------------------------------------------------------------
  Final couplings
------------------------------------------------------------

    Temperature 1:

    ------------------------------------
    Conf1  Conf2    Avg.   Exp.  Diff.
    ------------------------------------
     7.47   2.21   4.48   4.48    0.00
     5.31   5.29   5.30   5.30    0.00
     0.81   9.05   5.50   5.50    0.00
    ------------------------------------
    RMSD :     0.00 Hz


------------------------------------------------------------
  ERROR ANALYSIS
------------------------------------------------------------

  i       dP1/dPi     dP2/dPi     dP3/dPi     dP4/dPi     dP5/dPi    dRMSD/dPi
  1         1.000       0.258       0.510      -0.572       0.130      -0.000
  2        -2.615       1.000      -0.487      -1.127       0.692      -0.000
  3         4.238      -2.681       1.000       2.197      -1.283      -0.000
  4        -0.990      -0.697      -0.633       1.000      -0.371      -0.000
  5         0.875       2.444       2.116      -4.800       1.000      -0.000

  i       dP1/dCi     dP2/dCi     dP3/dCi     dP4/dCi     dP5/dCi    dRMSD/dCi
  1      -49.798       7.518       3.679     -25.782      20.703      -0.000
  2      -43.109      26.474      14.908     -18.933      23.287      -0.000
  3     -117.997       2.186    -112.890     -10.375      30.681      -0.000

------------------------------------------------------------
  RMSD               :       0.00 Hz
  Total time         :       0.84 s
------------------------------------------------------------
```

6.3.3 2-Methoxy-Uridine

```
% Electronegativities from ribose template

////////////////////////////////////////////////////////////

  START Pseudorotational calculation

////////////////////////////////////////////////////////////
////////////////////////////////////////////////////////////


Local minimum possible. Constraints satisfied.

No active inequalities.


------------------------------------------------------------
  Optimized parameters
------------------------------------------------------------

   Conformation 1
    P     :   34.355
    Phi_m :   35.492

   Conformation 2
    P     :  162.726
    Phi_m :   30.917

   Temperature Coefficients
    %Conformation1  :   60.233

------------------------------------------------------------
  Endocyclic torsion angles
------------------------------------------------------------
```

	Conf.1	Conf.2
Phi0:	29.646	-29.463
Phi1:	-35.545	18.187
Phi2:	27.867	0.036
Phi3:	-9.545	-18.245
Phi4:	-12.423	29.485

```
------------------------------------------------------------
 Final couplings
------------------------------------------------------------

  Temperature 1:

   ------------------------------------
   Conf1  Conf2    Avg.   Exp.  Diff.
   ------------------------------------
    1.43   7.70   3.92   3.92   -0.00
    5.08   5.38   5.20   5.20    0.00
    8.75   0.98   5.66   5.66   -0.00
   ------------------------------------
   RMSD :     0.00 Hz


------------------------------------------------------------
 ERROR ANALYSIS
------------------------------------------------------------

 i       dP1/dPi     dP2/dPi     dP3/dPi     dP4/dPi     dP5/dPi    dRMSD/dPi
 1         1.000       0.077       2.257       0.175       0.140      -0.000
 2         0.583       1.000       0.144      -0.933      -0.531      -0.000
 3         0.505       0.128       1.000      -0.031       0.014      -0.000
 4        -0.386      -0.827      -0.242       1.000       0.528      -0.000
 5        -0.088      -1.393       0.585       1.899       1.000      -0.000

 i       dP1/dCi     dP2/dCi     dP3/dCi     dP4/dCi     dP5/dCi    dRMSD/dCi
 1        -2.051      -3.778     -31.546      13.493      -1.151      -0.000
 2        12.264      -0.758     -29.187     -18.272     -10.879      -0.000
 3         2.090      -7.188     -27.318      22.130      19.079      -0.000

------------------------------------------------------------
 RMSD                :       0.00 Hz
 Total time          :       0.63 s
------------------------------------------------------------
```

6.3.4 2'-Fluoro-Deoxyuridine

```
%Electronegativity position 9 changed to 1.37, rest of values taken from rU template

////////////////////////////////////////////////////////////

  START Pseudorotational calculation

////////////////////////////////////////////////////////////
////////////////////////////////////////////////////////////


Local minimum possible. Constraints satisfied.

No active inequalities.


------------------------------------------------------------
  Optimized parameters
------------------------------------------------------------

   Conformation 1
    P     :   36.089
    Phi_m :   34.186

   Conformation 2
    P     :   38.134
    Phi_m :   34.176

   Temperature Coefficients
    %Conformation1  :   96.888

------------------------------------------------------------
  Endocyclic torsion angles
------------------------------------------------------------
```

	Conf.1	Conf.2
Phi0:	27.968	27.945
Phi1:	-34.265	-34.256
Phi2:	27.474	27.482
Phi3:	-10.189	-10.211
Phi4:	-10.988	-10.960

```
------------------------------------------------------------
 Final couplings
------------------------------------------------------------

  Temperature 1:

   ------------------------------------
   Conf1  Conf2    Avg.   Exp.  Diff.
   ------------------------------------
    1.41   1.41   1.41   1.38    0.03
    4.95   4.95   4.95   4.99   -0.05
    8.61   8.61   8.61   8.63   -0.02
   ------------------------------------
   RMSD :     0.03 Hz


------------------------------------------------------------
 ERROR ANALYSIS
------------------------------------------------------------

 i       dP1/dPi     dP2/dPi     dP3/dPi     dP4/dPi     dP5/dPi    dRMSD/dPi
 1         1.000       0.090      -0.015      -0.003     -18.504        0.000
 2         1.282       1.000      -0.013      -0.003     -18.504        0.000
 3         0.001       0.000       1.000      -0.292       0.594        0.000
 4         0.000      -0.000      -0.156       1.000       0.594       -0.000
 5         0.000       0.000       0.000       0.000       1.000        0.000

 i       dP1/dCi     dP2/dCi     dP3/dCi     dP4/dCi     dP5/dCi    dRMSD/dCi
 1        20.139      17.336     -15.031      36.082     -33.861       -0.034
 2         4.987      -2.483       4.526       1.499       5.519        0.233
 3         6.661       7.601       6.566      -7.616     -20.497        0.126

------------------------------------------------------------
 RMSD               :       0.03 Hz
 Total time         :       0.39 s
------------------------------------------------------------
```

6.3.5 2'-Methylthio-Deoxyuridine

```
%Electronegativity at position 9 changed to 0.785, literature results for -SR group;
%rest of values from ribose template

////////////////////////////////////////////////////////////

  START Pseudorotational calculation

////////////////////////////////////////////////////////////
////////////////////////////////////////////////////////////


Local minimum possible. Constraints satisfied.

No active inequalities.


------------------------------------------------------------
  Optimized parameters
------------------------------------------------------------

   Conformation 1
    P     :   50.661
    Phi_m :   16.656

   Conformation 2
    P     :  127.057
    Phi_m :   45.098

   Temperature Coefficients
    %Conformation1  :   23.584

------------------------------------------------------------
  Endocyclic torsion angles
------------------------------------------------------------
```

	Conf.1	Conf.2
Phi0:	-14.440	-26.741
Phi1:	6.756	0.209
Phi2:	3.509	26.402
Phi3:	-12.433	-42.929
Phi4:	16.609	43.058

```
------------------------------------------------------------
 Final couplings
------------------------------------------------------------

  Temperature 1:

   ------------------------------------
   Conf1  Conf2    Avg.   Exp.  Diff.
   ------------------------------------
    5.29   9.19   8.27   8.27    0.00
    6.31   5.63   5.79   5.79    0.00
    2.12   3.05   2.83   2.83   -0.00
   ------------------------------------
   RMSD :     0.00 Hz


------------------------------------------------------------
 ERROR ANALYSIS
------------------------------------------------------------

 i       dP1/dPi     dP2/dPi     dP3/dPi     dP4/dPi     dP5/dPi    dRMSD/dPi
 1         1.000       0.011      -0.083       0.058       0.004       0.000
 2         0.094       1.000       0.248      -0.397      -0.850       0.000
 3       -15.075       2.266       1.000      -1.537      -2.494       0.000
 4         2.079      -1.816      -1.040       1.000       2.002       0.000
 5        -0.108      -0.922      -0.319       0.451       1.000       0.000

 i       dP1/dCi     dP2/dCi     dP3/dCi     dP4/dCi     dP5/dCi    dRMSD/dCi
 1        19.438      30.222       3.961      -9.581     -47.169       0.024
 2        -9.868     -13.068       3.034     -18.964     -47.169       0.206
 3        32.779      20.861      -7.126       9.082      -7.419       0.000

------------------------------------------------------------
 RMSD               :       0.00 Hz
 Total time         :       0.56 s
------------------------------------------------------------
```

6.3.6 2'-Selenomethyl-Deoxyuridine

```
%Electronegativity for 2' from Manual, rest of values from ribose template

////////////////////////////////////////////////////////////

  START Pseudorotational calculation

////////////////////////////////////////////////////////////
////////////////////////////////////////////////////////////


Local minimum possible. Constraints satisfied.

No active inequalities.


------------------------------------------------------------
  Optimized parameters
------------------------------------------------------------

   Conformation 1
    P     :    7.165
    Phi_m :   20.308

   Conformation 2
    P     :  134.686
    Phi_m :   40.370

   Temperature Coefficients
    %Conformation1  :   17.768

------------------------------------------------------------
  Endocyclic torsion angles
------------------------------------------------------------
```

	Conf.1	Conf.2
Phi0:	20.231	-25.056
Phi1:	-17.695	5.558
Phi2:	8.401	19.064
Phi3:	4.103	-36.404
Phi4:	-15.039	39.839

```
------------------------------------------------------------
 Final couplings
------------------------------------------------------------

  Temperature 1:

   ------------------------------------
   Conf1  Conf2    Avg.   Exp.  Diff.
   ------------------------------------
    1.62  10.17   8.65   8.65   -0.00
    7.00   5.44   5.72   5.72    0.00
    6.05   2.17   2.86   2.86   -0.00
   ------------------------------------
   RMSD :     0.00 Hz


------------------------------------------------------------
 ERROR ANALYSIS
------------------------------------------------------------

 i       dP1/dPi     dP2/dPi     dP3/dPi     dP4/dPi     dP5/dPi    dRMSD/dPi
 1         1.000       0.026       0.095      -0.062       0.004       0.000
 2         0.247       1.000       0.181      -0.416      -0.280      -0.000
 3         6.647       4.577       1.000      -1.557      -0.903       0.000
 4        -0.533      -1.967      -0.497       1.000       0.616       0.000
 5         2.274      -2.911      -0.592       1.435       1.000      -0.000

 i       dP1/dCi     dP2/dCi     dP3/dCi     dP4/dCi     dP5/dCi    dRMSD/dCi
 1       -19.912     -59.176     -18.903      25.203      20.176       0.000
 2         6.555     -36.335     -11.634      -7.149      -1.509      -0.000
 3        39.923      14.556      -4.233       1.885       2.324       0.000

------------------------------------------------------------
 RMSD               :       0.00 Hz
 Total time         :       0.67 s
------------------------------------------------------------
```

PART TWO

EXTRACELLULAR RECOMBINANT PROTEIN PRODUCTION: CURRENT SYSTEMS AND APPLICATIONS

7 INTRODUCTION

7.1 The Central Dogma of Molecular Biology

The complexity of the living world is governed, maintained and perpetuated by a fairly simple process that is conserved from bacteria to humans. This flow of information is so critical to the function of biological processes and the understanding of the concepts involved that it is designated the central dogma of molecular biology.[1] The genetic code, the series of chemical bytes that contains the information needed for construction of a living entity, consists of DNA. This biological polymer is characterized by a phosphate-linked chain of sugars attached to a planar moiety called a nucleic base. This base is the information carrier and a group of three bases encodes a specific amino acid. A DNA polymer is very long, contained as a chromosome within an organism's genome. These chromosomes are safe-guarded during normal cellular operation and replicated when a cell is dividing during growth. Segments of the chromosome, called genes, are accessed and transcribed into shorter, similar oligomers consisting of RNA. These messenger RNAs contain the blueprint for a protein as encoded by the DNA and through the action of ribosomes (large protein/ RNA complexes) the signal is translated into a second type of biopolymer, a protein. A simplified picture of these processes is shown in Figure 7.1.

There are viruses whose replication cycle is dependent on reverse transcription of RNA to DNA or the replication of RNA, and some eukaryotic processes such as telomere elongation or RNA interference use similar mechanisms.[2-3] The general case, however, of DNA transcription to RNA then translation to a protein is the most common sequence of events.

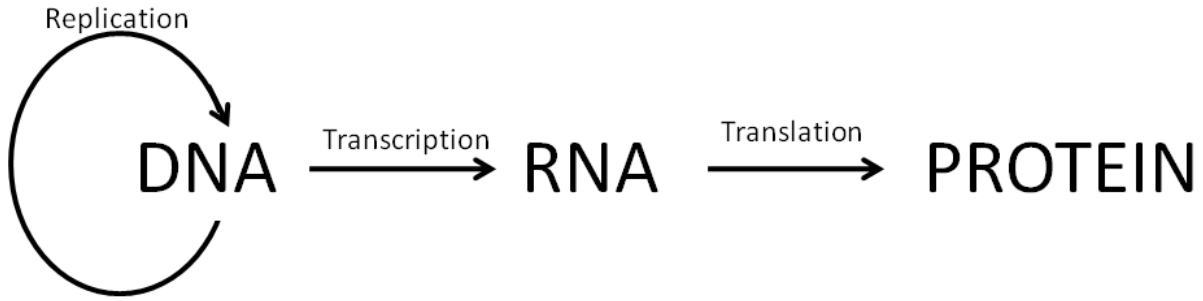


Figure 7.1 Central Dogma of Molecular Biology

7.2 Protein Chemistry and Structure

Perhaps the most awe-inspiring concept in biochemistry is the ability of roughly 20 building blocks to combine to form structures that can catalyze reactions, bind substrates, recognize threats, and transport nutrients with such specificity and diversity. Proteins are known as the workhorses of a cell, building and degrading all major components while transporting and proofreading unfinished products under highly-tuned regulation. The populations and chemical properties of the amino acid building blocks within the peptide-bond linked polymers are that which gives rise to this diversity of form and function.

The amino acids themselves are built on the same form, consisting of a carboxylic acid and an amine linked to a carbon atom that also contains an additional functional group, the identity of which determines the identity of the amino acid. The stereochemistry of the α-carbon (the atom to which the amine, carboxylic acid, and functional group are attached) can be either D- or L- in the convention of chirality, but interestingly only L-amino acids are utilized in biological systems for protein synthesis. Proteins, when compared to other biological polymers, are much more diverse in

function because the side chains of these amino acid monomers can range from acidic to basic, polar to non-polar, and aromatic to aliphatic. In a cell, functional proteins are able to happily perform their roles based on this assortment of choices for building blocks.

The proper arrangement of these building blocks has a large effect on the function of a protein, and naturally on the overall structure. Many forces play in the folding of a long polypeptide into a functional protein, mainly hydrophobic interactions directing non-polar residues away from the water solvent and align the side chains in a relatively high density.[4] Secondary structural elements of a protein are often characterized by patterns in the sequence, as an alternation between polar and non-polar amino acids can contribute to the formation of an α-helix or β-sheet depending on the pattern. These secondary structure components put the appropriate amino acids that are not necessarily close in sequence near to each other in space allowing for a protein to be active. In addition, in eukaryotic proteins, a degree of post-translational modification (e.g. methylation, acetylation, glycosylation) is necessary for correct function.

7.3 Diversity in Environment and Function

Life inhabits a great majority of the surface of the earth, and because of these drastically different environments there is an even wider range of metabolic diversity. Phototrophs, nitrogen-fixers, hydrogen sulfide oxidizers, anaerobes, and methanophiles all use specific proteins to accomplish the continuous act of using energy. These proteins are usually specific in their action as well as optimized for the conditions.

The action of a protein is usually concentrated in a specific subspace of a cell or tissue. This is easy to accept when considering the ineffectiveness of a transcription factor embedded into the plasma membrane or cardiac myosin fibers forming in the pituitary gland. The inherent need for a tightly regulated transport system is clear, and nature utilizes two broad categories of protein targeting: Co-translational and post-translational, which are visualized in Figure 7.2. Initiation of translation occurs but when the nascent peptide emerges from the ribosomal tunnel, the pathways diverge. The co-translational transport pathway is initiated like normal protein synthesis but is paused while a chaperone directs the ribosome to the cell membrane and the protein is inserted into the cell membrane or the periplasm. The action of membrane translocation is usually coupled with GTP hydrolysis.[5] In contrast, as the name implies, the post-translational mechanism occurs after translation. A different signal sequence is contained within a post-translationally transported protein and is directed to one of many locations within the cell such as the periplasm in gram negative bacteria or various organelles within eukaryotes. An interesting correlation between the two processes is the hallmark of post-translational transport: a fairly conserved set of ATP binding components to shuttle proteins across membranes like SecA in prokaryotes and HSP70, HSP40 and ASNA1 in eukaryotes.[5] A main feature of the transport factors in both cases is centered on assuring the continuance of a translation competent state, in the sense of assurance that the fresh peptide chains do not fold too early or aggregate with other cellular components. Proteins that do not have a signal sequence fold into their active states and usually remain in the cytoplasm.

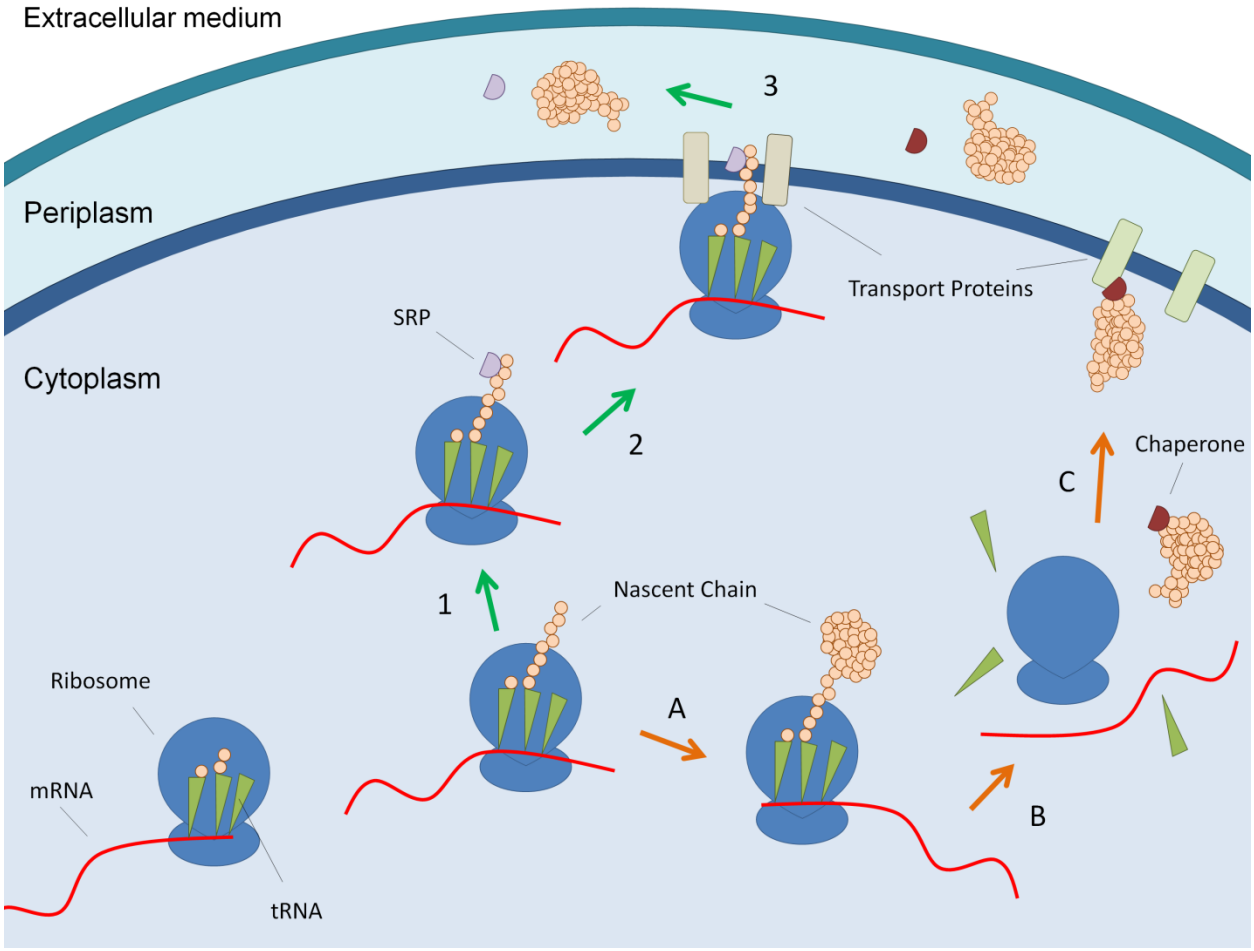


Figure 7.2 Co and Post-Translational Secretion Mechanisms. Co-translational transport begins with a pause in translation (1) and direction of the ribosome to the cell membrane (2). Translation continues and the peptide is directed into the periplasm (3) or the cell membrane itself. Post-translational Protein Transport does not pause translation (A) and the signal sequence is recognized after translation occurs (B), with the newly synthesized peptide being exported across the membrane upon sequence recognition (C).

7.4 Commercial Protein Production and the Perspective of this Manuscript

The diversity of protein function inevitably leads to a large economic value for certain proteins that can perform a specific task. Digestive enzymes are critical for paper, detergent, food, and fuel industries. Antibodies, hormones, and biopolymer building blocks are relevant in medicine. As the world becomes more advanced

technologically, the demand for these and other biologically based products will only increase. Industries use different organisms to make the intended products, often in a recombinant fashion where the protein is produced in a species different from where it is naturally found, and current protocols have been effective in the manufacturing of acceptable amounts of various products. Still, each method is a culmination of many attempts to find optimum conditions at a scale that is economically suitable.

There has been a push more recently to try and develop more effective protein production systems,[6-7] where purification can be simpler while increasing yields without significantly altering the growth cycles of the cultures. One of the more realistic ways of achieving these goals is to use secretion platforms to transport proteins of interest outside of the cell, away from the majority of cellular components that can contaminate and degrade the recombinant products. If the protein is secreted, then the purification process is greatly simplified by not separating out the entire proteome within the cell, as is the case with intracellular expression and lysis. The direction of proteins out of a cell allows for them to be drained away in a reactor during the cycling of nutrients and purified afterwards in fewer steps. This approach is also appealing because the cultured cells would not need to be lysed in order to retrieve the products and could remain in an induced production state.

This paper is an attempt to distill relevancy and perspective from the vast amount of information available about protein production and secretion, including current and future applications, while directing interested parties towards further research.

8 PROKARYOTIC PROTEIN EXPORT

8.1 Prokaryotic Cell Biology

Prokaryotes (bacteria, archaea) are the simplest forms of life while genetically and metabolically the most diverse. These single-celled creatures are robust and highly specialized, which contributes to their diversity. A generalized model of both domains depicts the prokaryotic cell as a living soap bubble. The DNA, RNA, and proteins within the cell are all exposed to each other in a mixture of metabolites and salt. The genetic material is constantly being transcribed and translated into its encoded products surrounded by a lipid bilayer (at least one, and possibly by a cell wall) that keeps the cell's contents separate from the extracellular medium from which sustenance and stimuli are sourced. The cell membrane is spotted with proteins that form channels for the maintenance of cellular homeostasis and receptors for a variety of substrates that the cell deems worthy of engulfing or avoiding. The ribosomes that translate the mRNA into proteins are free in the cytoplasm, which allows for quick responses to stimuli, but still requires a system of organization for controlling this process.

The vast array of different metabolic pathways utilized by bacteria is far beyond the scope of this paper, but it is worth knowing that the mechanism for energy consumption that a prokaryote employs because of the high level of regulation involved in different periods of metabolic functioning. Times of starvation cause the activation of certain genes and the repression of others, while times of abundant resources have the opposite effect. Normal functioning requires regulation as well. Prokaryotes do not have compartmentalized organelles but do contain localizing signal sequences that direct membrane or periplasm proteins to their appropriate places. Extracellular proteins are

directed outside by several secretion systems, but the degree to which this happens depends on the species. For example, heterotrophic organisms would be expected to secrete digestive enzymes to obtain energy containing compounds, while phototrophic microbes would not need to export proteins as their energy comes from light and their carbon source is atmospheric CO_2.

8.2 Prokaryotic Secretion Mechanisms

8.2.1 Type I Secretion

Gram negative bacteria have a harder time exporting protein than gram positive because their additional LPS-containing outer membrane is an extra step in the process. The type I secretion system (T1SS) is one mechanism gram negative bacteria use to overcome this barrier. T1SS is composed of three proteins located in the cell membranes, each of which is required for secretion. T1SS allows for the secretion of proteins of various sizes and functions from the cytoplasm to the extracellular medium in a single step. The process begins with recognition of a C-terminal, glycine-rich signal sequence by an ATP-binding cassette (ABC) protein. The exact sequence of the signal varies between target proteins and bacteria species. A general consensus is a repeat of GGXGXDXXX, from a few to more than 50 repeats depending on the protein and species. This segment is almost always acidic and can bind calcium specifically. The ABC protein is bound to a membrane fusion protein (MFP) which in turn is linked to an outer membrane protein (OMP). This complex modulates the opening and closing of a tunnel through both membranes allowing the protein to transverse, and all three of the components are required for secretion to occur. This mechanism is reviewed in a step-wise process by Delepelaire[8] quite effectively. A general scheme of the T1SS is shown

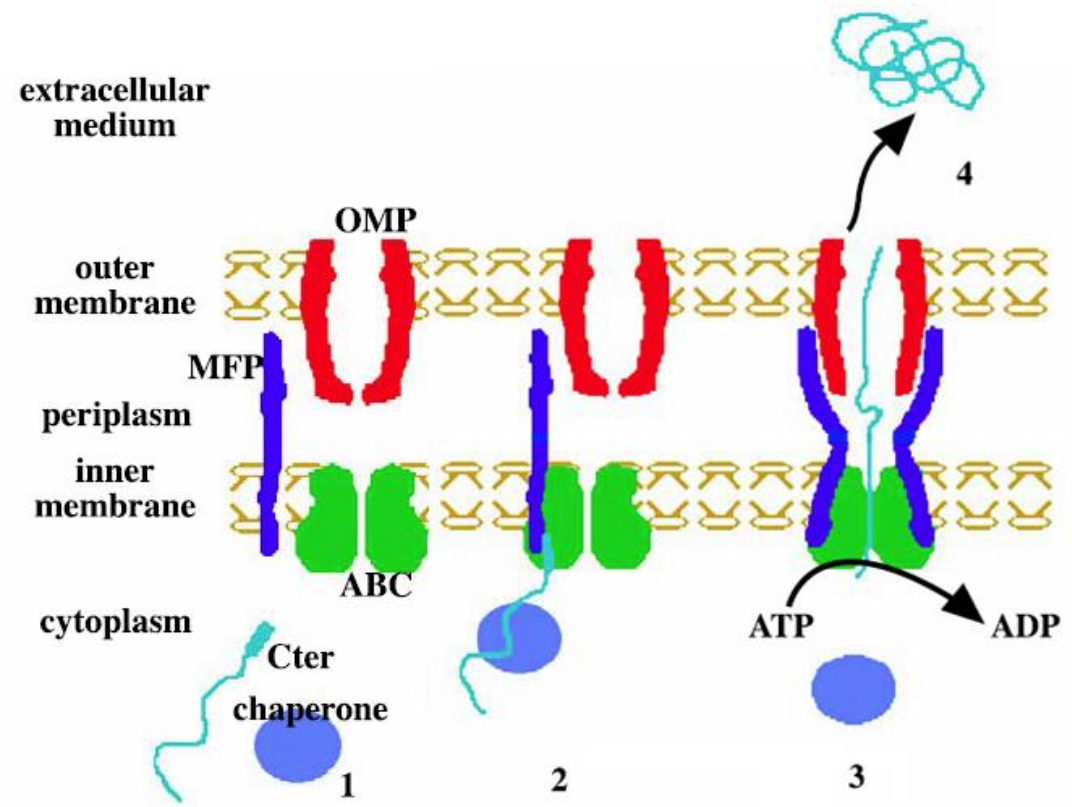


Figure 8.1 Mechanism of protein export by T1SS. (1) C-terminal signal sequence is recognized by the ATP-binding cassette (ABC) while chaperone proteins keep the polypeptide unfolded. (2) ABC forms complex with the membrane fusion protein (MFP) and outer membrane protein (OMP) while binding ATP. (3) ATP hydrolysis drives the opening of the channel and the secretion of the protein. Figure adapted from Delepelaire[8].

in Figure 8.1. One of the more widely studied examples of a T1SS is the hemolysin A pathway, a protein secreted by pathogenic bacteria such as *E. coli* and *S. aureus* that lyses red blood cells and binds free iron. Hemolysin A (HlyA) is the name of the 46-60 amino acid signal sequence in *E. coli*, and it is recognized by HlyB and HlyD before being exported by TolC,[9] the OMP for the HlyA pathway and many others. This pathway is popular because of its simplicity and capability of transporting large (up to 800 kDa) proteins through its 20-30 Å pore, but is naturally only capable of yields in the 10mg/ L range.[10]

T1SSs have been demonstrated to export a wide range of proteases, lipases and adhesins in wild-type gram negative bacteria, as well as a reasonable amount of recombinant fusion proteins. The signal sequence itself has been examined using a systems biology approach, basically assuming DNA encoding glycine rich residues upstream of stop codons correlate to T1S signal sequences.[9] The beauty of the T1SS is that it can send a peptide across both cellular membranes in a single step,[11] however undesired translation rates must be kept to a minimum in order to avoid clogging the machinery.[12]

8.2.2 Type II Secretion

The second type of secretion mechanism characterized in gram negative bacteria, the aptly named Type II secretion system (T2SS), proceeds via a two step process, in contrast to T1S. The first step is translocation across the inner membrane and an amount of time residing in the periplasmic space. The second step is secretion from the periplasm to the exterior of the cell. The molecular machinery that executes the first step of the T2S process is a fairly complicated structure. A review of 12 different genera by Cianciotto elegantly compiles the conserved elements and describes their functions.[13] There are three related mechanisms by which T2S can be executed: The Sec-dependent pathway, a signal recognition particle (SRP) pathway, or the twin-arginine translocase (TAT) pathway.[9] Each of these pathways has their own versions of pseudopilins, which facilitate the binding of the ATPase attachment to the inner membrane and form the pore through the membrane. The pathways also include the secretin pore, which forms upon the congregation of secretin monomers in the outer membrane, and a protein involved in substrate recognition.[13]

The Sec pathway is a post-translational secretion system. This route is directed by a cleavable N-terminal signal sequence that can vary with respect to the protein being secreted or the species involved. The protein led by the signal sequence is delivered to the periplasm or outer membrane and then the signal is cleaved.[14] Because this is a post-translational mechanism, there are chaperone proteins that are known to interact with the unfolded region of the peptide to discourage aggregation and to maintain the full secretory proficiency. SecB is the most common chaperone in the Sec pathway, and it is interesting to note that it does not bind to the signal sequence itself; rather it holds extensive unfolded segments of the substrate with little specificity. The only known protein SecB binds with high affinity is SecA, which is associated with the SecYEG export channel and is believed to assist in proper delivery to the pore.[5] A few other cytosolic chaperones that are related to heat-shock proteins of the HSP70 family have been reviewed by Cross et al.[5] and have been well characterized as intermediates in protein folding and cell stress. These proteins have been shown to help recover some protein secretion in SecA/SecB knockout strains.[15] These particular HSP70 proteins, DnaJ and DnaK are ubiquitous proteins in gram-negative bacteria and seem to be involved in Sec-independent secretion pathways as well.[16] Another interesting finding regarding the Sec-pathway is the dependence on the amino acid flanking the signal sequence for efficient protein export. The study by Kaderbhai et al. indicates that smaller, more hydrophilic amino acids are ideal for the flanking residue.[17]

The SRP pathway is a co-translational pathway facilitated by an N-terminal, highly hydrophobic sequence, similar to SecB, and actually uses the same translocation machinery (SecYEB complex) upon binding to SRP.[18] The hydrophobicity of the N-

terminal sequence determines which cofactor will bind and direct export. One can engineer hydrophobicity into the sequence and recruit the SRP, which tends to help when proteins fold too rapidly for SecB transport. A downside to this technique as a protein production avenue is that the proteins accumulate in the periplasm and do not reach significant levels outside of the cell. Ways to overcome this problem are discussed in Section 11.4.

The twin-arginine translocase (TAT) pathway is independent of Sec machinery and ATP hydrolysis, and is characterized by the highly conserved double arginine segment of its signal sequence, actually using the intermembrane's proton motive force to drive export. What is interesting about the TAT pathway is that it can transport fully folded proteins across the inner membrane, including ones with redox cofactors. *In silico* predictions suggest that roughly 20% of secretion in *Streptomyces* is dependent on the TAT process.[19] The TAT pathway has been used to moderate success for the production of extracellular recombinant proteins, and there have been reports of increased protein production upon the overexpression of TAT translocation elements.[20-21] This is promising considering the fact that wild-type TAT dependent secretion is fairly low in both *E. coli* and *B. subtilis*.[22-23] A comparison between the different T2SSs is given in Figure 8.2. The evidence of the role of T2SS machinery in the pathogenesis of many bacteria is present in systems biology studies of the genomes of fish, mammal, and plant pathogens. This data is drawn from analysis of potential secretion signal sequences, the degradative nature of known T2SS dependent proteins, the fact that some mutations in T2SS genes can lower virulence in some relevant plant and animal

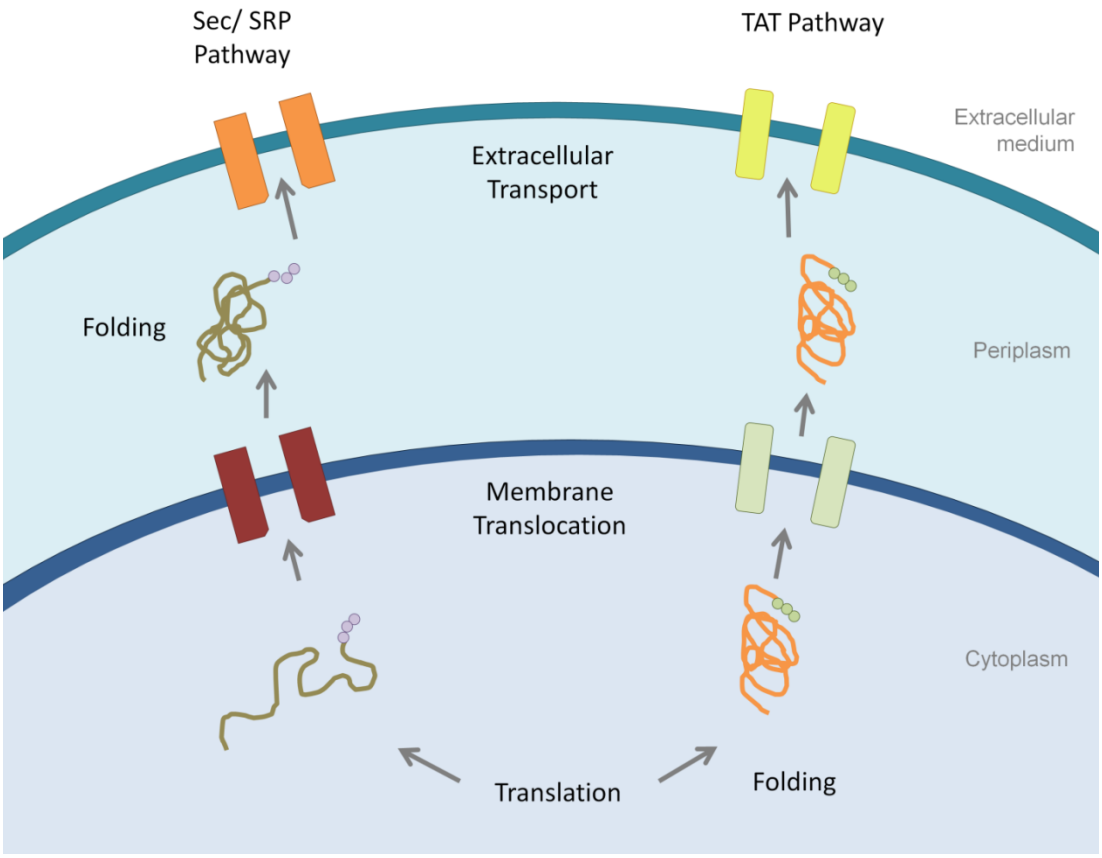


Figure 8.2 Comparison of T2SS mechanisms

disease models, and the few instances where individual T2 exoenzymes have been shown to contribute to pathogenicity,[13] namely cholera toxin.[24]

However, there are several instances of non-pathogenic bacteria utilizing a T2SS: *A.calcoaceticus* secretes dodecane degradative enzymes,[25] *A.alcaligenes* secretes lipases,[26] *P.putida* exports manganese and iron reducing enzymes[27] similar to *S. oneidensis*,[28] and *G. diazotrophicus* secretes a levansucrase.[29]

8.2.3 Type III Secretion

The type III secretion system family (T3SS) of protein complexes is based on a complicated macromolecular machine consisting of a secretion apparatus, chaperones, the secreted proteins themselves, and cytoplasmic regulators. Flagellar proteins are

often dependent on T3SS machinery, and a large portion of the proteins secreted are known to be involved in the virulence of pathogens that allow them to colonize a niche within the host, evade immune system detection, enter or exit host cells, or obtain nutrition.[30]

The T3SS is built across the membranes and protrudes outside of the cell. It acts as a sensor for host contact. The assembly of the machinery begins with the formation of a basal body complex which spans the inner and outer bacterial membrane and acts as the scaffolding for needle subunits and needle length control proteins to build the injection apparatus. The needle subunits form a hollow extracellular structure and upon reaching the ideal length, a substrate switch is activated and the tip proteins form the completed assembly and secretion is halted until host cell contact occurs. The pore subunits are secreted when the pathogen makes host cell contact. They insert into the host membrane to form another pore and initiate the final substrate switch to begin secretion of virulence factors into the host cell. The mechanism of this complicated process is regulated at transcriptional, post-transcriptional, and translational levels and is dependent on the conformational switching of the components and the actual secretion of the building blocks.[31-32] Deane et al. have reviewed this process in detail,[30] and Figure 8.3 shows a model of a mature T3S apparatus. T3S also directs proteins to its translocation process with a signal sequence. There has been little agreement regarding the nature of the signal because there is neither a cleavable sequence nor a recognized consensus in sequence. The exact mechanism of T3S regulation is still being debated, as evidence shows that the mRNA and the peptide signal sequence are critical in some cases while unnecessary in others.[33] However, recently, a minimal, 22-

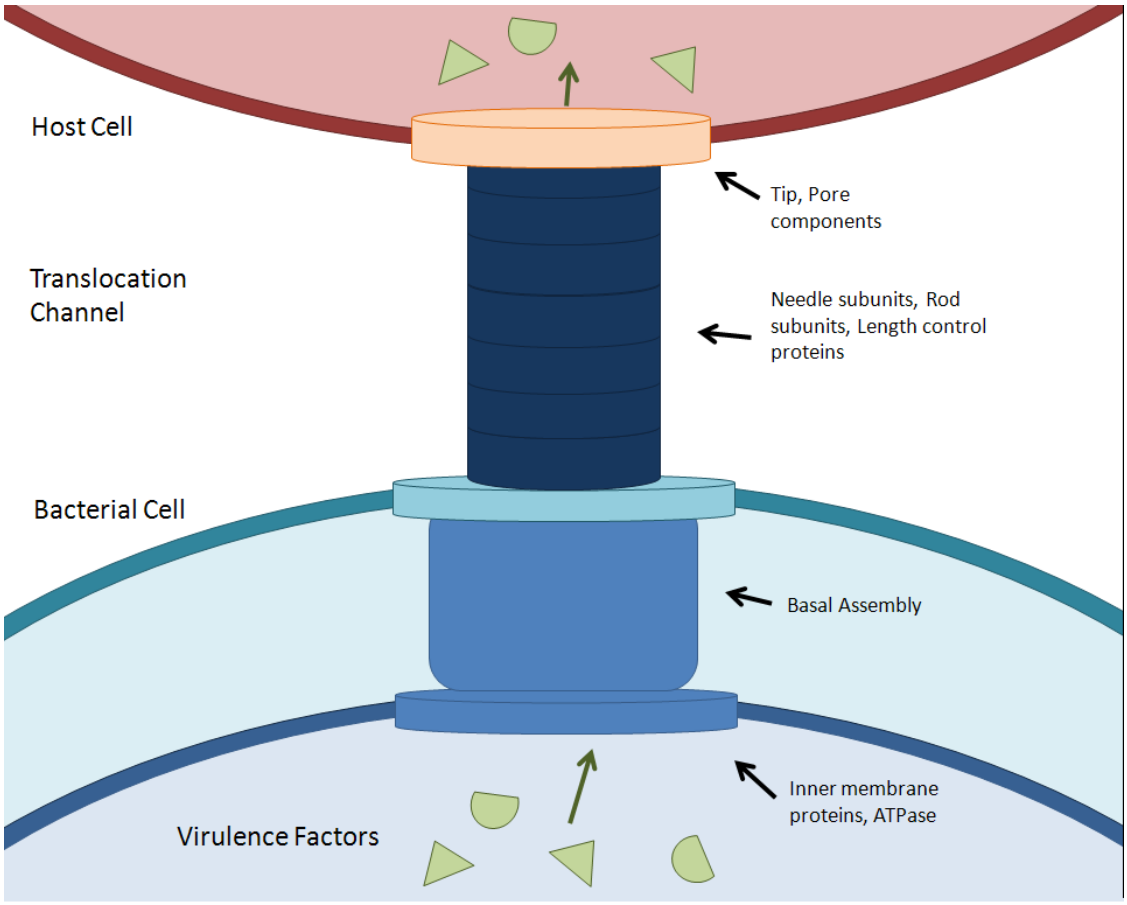


Figure 8.3 Illustration of the T3SS/ Host Cell Conjugate Apparatus

residue sequence at the N-terminus of a flagellin molecule has been identified to contain all essential information to direct the protein to the T3SS.

Furthermore, the T3S export channel does not seem to be highly specific for the flagellar proteins, but is capable of exporting a variety of proteins.[33] A caveat to consider during the construction of T3S-dependent production systems is the necessity of chaperone binding sites as well as a proper signal sequence, as deletion of either has been shown to diminish secretion.[34] The pore itself is narrow (25 – 30 Å) and so proteins are expected to pass through in a mostly unfolded state,[35] which also puts a constraint on the applicability of certain recombinant protein production. However, the

core components of the T3SS are conserved between species,[36] and data suggests that there are conserved targeting mechanisms as well. T3SSs from one type of bacteria can in some cases export proteins from another bacterial species with the foreign signal sequence, and some small molecule inhibitors of T3S have been shown to be effective across many genera.[37] The consensus seems to be that the proton motive force (PMF) drives the translocation of peptides through the secretion apparatus in conjunction with ATP hydrolysis. The ATP hydrolysis only occurs when a chaperone protein comes in contact with an ATPase that is part of the inner membrane complex, and a conformational change is induced in the complex. The pH gradient is suggested at the carrying force, as opposed to an electronic potential difference facilitating motion, which means that T3S is similar to the Sec and TAT pathways in this regard.[38]

8.2.4 Types IV - VII

There are other types of secretion systems employed by prokaryotes for various functions, but they share similarities to the systems described above. Type IV secretion (T4S) is a sophisticated process, dependent on the formation of a large secretion platform, and is related to conjugation transfer machines that transport nucleic acid-protein complexes between two cells. This transfer of genetic material and protein can occur between two species of bacteria or between bacterial pathogens and their hosts.[39-40] The consensus is that the process transports its cargo across both membranes without an extended visit to the periplasm once the transport machinery has assembled. The models of action proposed are designated as the channel model and the piston model. As their name implies, in the first mechanism the pilus structure opens a channel and allows the proper substrate to pass through the membranes, while

the second mechanism has the pilus acting like a molecular piston, pushing the substrate through the membranes.[41]

Type V secretion (T5S) has an interesting pathway. The two most studied variants of the T5SS are named the autotransporter complex and Twin-partner secretion, and both consist of a family of virulence factors that use the Sec pathway to enter the periplasmic space and then, via a conserved C-terminal β-domain that inserts into the outer membrane, serves to secrete and cleave the N-terminal passenger domain.[42] The proposed mechanisms for this action, which are mediated by periplasmic chaperone proteins,[43-44] can vary on the final cleavage of the passenger domain from the translocation domain. The autotransporter mechanism is described as the hairpin model and suggests that the C-terminal linker section connecting the β-domain to the secreted domain enters the translocation channel first and then threads the remaining portion of the protein through, and the folding of this secreted domain into an α-helix pulls the peptide through the channel. Upon the full transport and folding of the peptide, it can remain attached to the translocation element or can be cleaved by extracellular proteases. In the twin-partner secretion pathway, the translocation element and the secreted protein are not fused in a propeptide-like product, but are usually encoded within the same operon. Both types are directed to the periplasm through a T2SS pathway, and after the translocation element inserts into the outer membrane, it recognizes its partner's N-terminal sequence and proceeds to export it out of the cell.[42]

Type VI secretion (T6S) is another secretion mechanism of gram negative bacteria, originally thought to be a part of T4S, but seems to only share chaperones in certain cases.[45] After further study, the T6SS has been shown to function in some

capacity in quorum sensing,[46-47] as well as containing components that are host specific.[48] The data suggests that T6S is important in deciding the fate of various bacteria in polymicrobial environments, and it is known that the machinery resembles the needle apparatus common to other secretion systems, but the mechanisms and regulation of the T6SSs are still largely unknown.[49]

Type VII secretion (T7S) has been recently identified as a unique function of mycobacteria, allowing them to achieve virulence as well as cell to cell communication through their highly hydrophobic mycolic acid membrane.[50] The process is also referred to as the ESX pathway, and consists of a large set of proteins contained in a particular gene cluster. The process is very complicated, involving many transmembrane domains involved in recognition and guidance of apparatus assembly, some of which may be involved with some T4SS components, though the debate is still occurring. Although noted to be a critical part of certain pathogenic life cycles, such as *Mycobacterium tuberculosis*, this system has not been studied to the extent as some of the previous systems and the scope of its coverage will be limited here, though Simeone et al. have reviewed the relationship between mycobacterial T7S with host infection quite effectively.[51]

Each of the secretion pathways discussed in this section has value for the microbes that employ them, however, they have not been characterized to the point of exhibiting potential for extracellular recombinant protein production. Figure 8.4 is included to highlight the similarities and differences between each type of secretion systems found in prokaryotes.[52]

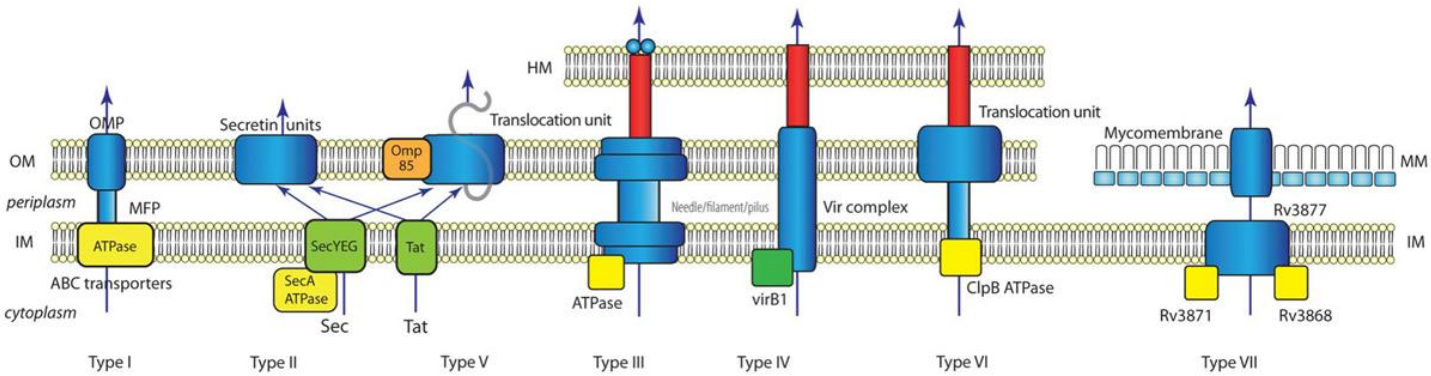


Figure 8.4 Overview of Known Bacterial Secretion Systems. Abbreviations are as follows: IM – Inner Membrane, OM – Outer Membrane, OMP – Outer Membrane Protein, MFP – Membrane Fusion Protein, HM – Host Membrane, MM - Mycomembrane. Figure taken without permission from Tseng et al., 2009.[52]

9 EUKARYOTIC PROTEIN EXPORT

9.1 Basic Eukaryotic Cell Biology

The central dogma of molecular biology holds true when discussing eukaryotes as well as prokaryotes, the main difference between these two being a level of complexity in physiology. Prokaryotes are by far more diverse metabolically, but eukaryotes are diverse in ways visible every day. All multicellular life is contained in the Eukaryota domain, and they are characterized by a series of intracellular organelles, each with their own membrane and self-contained processes. A sketch of the differences between prokaryotic and eukaryotic cells is shown in Figure 9.1.[53] The most definitive is the nucleus, which contains the genetic material of the cell and is the location of replication, transcription and mRNA editing. The nuclear environment is designed to protect the DNA from oxidative damage or foreign entities, and also contains the machinery for DNA repair. The mRNA synthesized in the nucleus is transported to the cytoplasm where signal sequences send the transcript to the

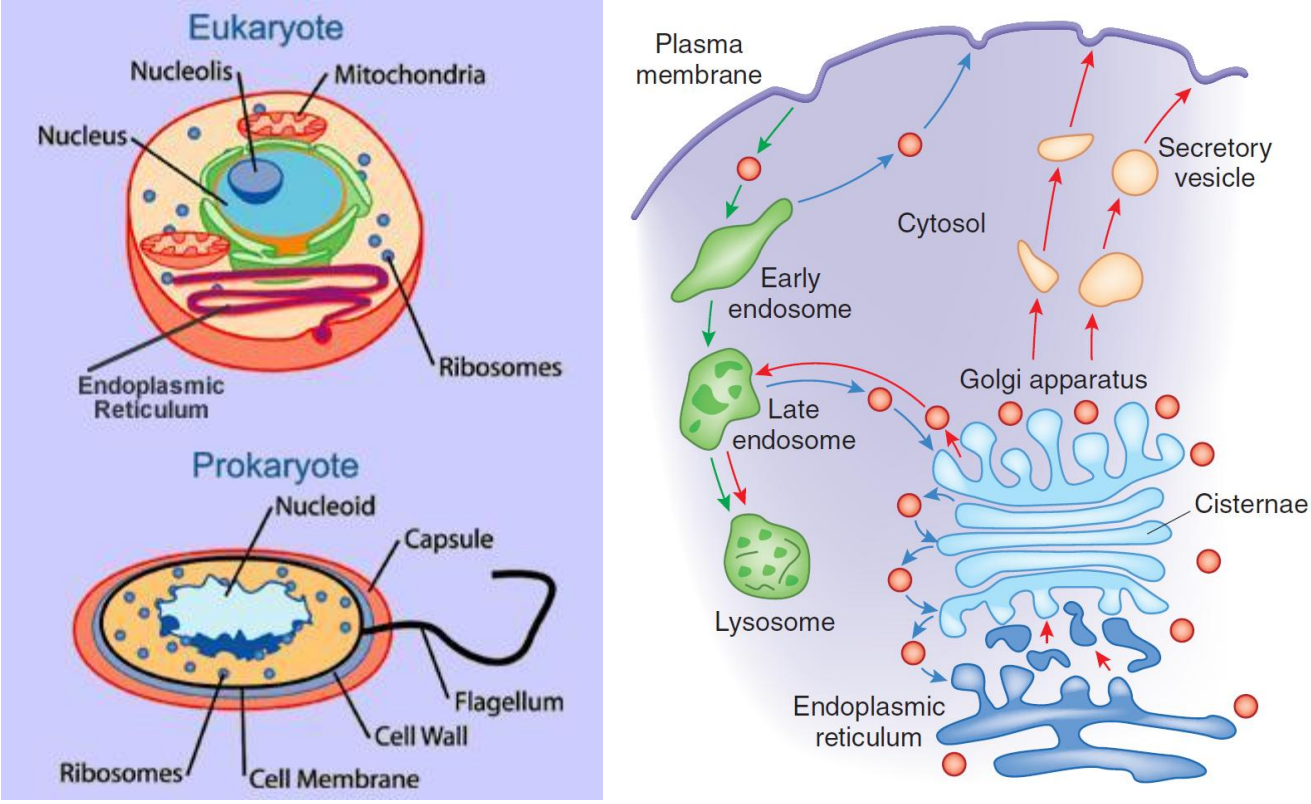


Figure 9.1 Overview of Eukaryotic Cell Biology. (Left) General comparison between eukaryotic and prokaryotic cell physiology. Picture from http://rst.gsfc.nasa.gov/Sect20/A12.html (Accessed April 10, 2011). (Right) Process of protein transport throughout a eukaryotic cell. Image from Xu and Esko, 2009.[53] Both images used without permission.

endoplasmic reticulum (ER) where they are translated into proteins by membrane-bound ribosomes. Eukaryotic protein structures are governed by the same intermolecular forces, and must be properly folded for function. The ER itself is a complex mixture of proteins that welcome incoming nascent peptides and assist in the folding process, consisting of a large quality control system. There are five main components of this system: Molecular chaperones that bind the nascent peptide until it can finish being translated before folding occurs, protein disulfide isomerases that reduce incorrectly formed disulfide bonds and allow the correct links to be connected, digestion machinery to process incorrectly folded proteins as part of the ER associated

protein degradation (ERAD), signal transduction pathways linked with this response, and a variety of post-translational modification enzymes involved with the final processing of synthesized proteins.[54]

Eukaryotic cells are generally much larger than prokaryotic cells, and because of the complexity of these systems and their compartmentalized physiology, the ability to recognize a signal sequence accurately and process it properly is an evolutionary function that eukaryotes have mastered.

9.2 Protein Sorting/ Targeting

Eukaryotes are inherently more complicated systems than prokaryotes because of the degree of compartmentalization within the cells. As stated above, transcription and RNA editing occur within the nucleus of a cell, while translation occurs in the cytoplasm, with or without being directed to the endoplasmic reticulum/ Golgi body network. The direction of proteins from the ER is mediated by signal sequences in the proteins themselves. Membrane bound proteins are sent to their respective organelles by traveling through the ER and Golgi apparatus until the formation of vesicles (small membrane-encapsulated micelle-like structures that carry proteins and metabolites through the cell with the help of cytoskeletal systems) transport the vesicle to its appropriate location imbedded into various cellular membranes. Other non-membrane proteins are sent inside the organelles themselves, some are exported out of the cell, while others remain inside the lumen of the ER or Golgi apparatus. The exception to this packaging of proteins for cell subspace delivery is the proteins of mitochondria and chloroplasts, a fraction of which are encoded by organelle DNA and transcribed and translated within the membrane of the organelle. The signaling sequences used in

eukaryotes can be located at the termini of the peptides or within the internal sequence, and many proteins require specific chaperones to fold correctly. In addition to the high degree of specific transport, many outer membrane proteins of eukaryotes have large amounts glycosylation sites that are used in recognition and communication between cells and must be added to the protein after translation. With all of these steps towards maturity, there are several proofreading systems in place that mark misfolded or mistranslated proteins with ubiquitin for degradation. Figure 9.2 gives a summary of targeting pathways in eukaryotic cells.[55]

Nevertheless, eukaryotic protein transport follows the same general pathways as prokaryotes: co-translational and post-translational. The co-translational is most effective for membrane bound proteins, as the incorporation of the nascent peptide directly into the membrane lowers the probability of misfolding as the large concentration of hydrophobic amino acids are not exposed to the aqueous environment of the lumen or cytoplasm and the energy of translation can drive the peptide forward. (See Section 8.2.2).The posttranscriptional pathway is used for nuclear, mitochondrial, chloroplastic, and peroxisomal proteins. The signal sequences used for different organelle targeting is so highly conserved that predicting the location of an newly discovered protein can usually be accomplished with just the sequence.[56]

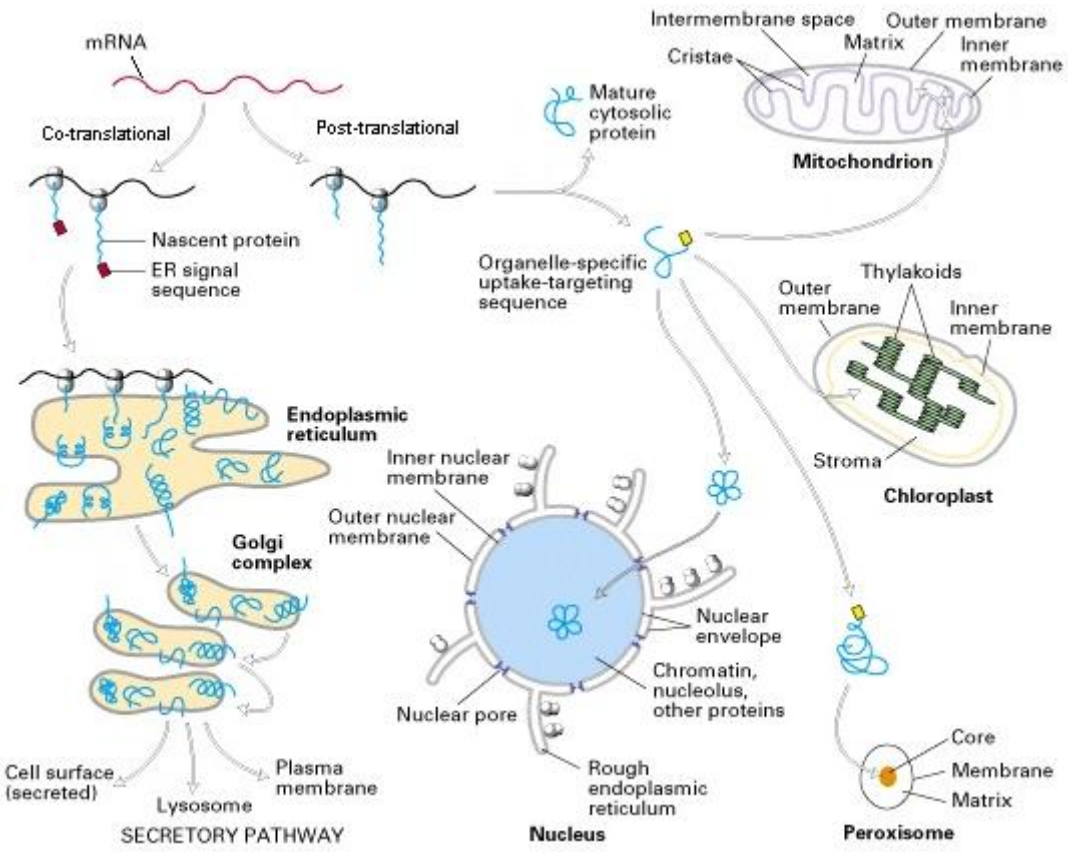


Figure 9.2 Summary of Eukaryotic Protein Targeting. Image adapted from Lodish et al. Copyright W.H. Freeman and Co.[55]

9.2.1 Signal Recognition Protein

The co-translational pathway is characterized by its own signal recognition protein (SRP) and dependence on GTP hydrolysis, similar in structure and function to the SRP of bacterial T2S.[57] The system contains the SRP, the SRP receptor, a GTPase that contacts the SRP, SRP receptor and ribosome, and an ER translocation pore. This SRP binds to the nascent peptide (N-terminal) and halts translation until the ribosome/ mRNA/ SRP complex is directed to the ER. Hydrolysis of GTP leads to conformational changes that reactivates translation and sends the peptide into the ER or the membrane.[5] The signal itself lacks any precise sequence arrangements, but recognition seems to be based on hydrophobicity of the peptide and possibly the secondary

structure (α-helical structures are formed in the output channel of the ribosome). Even though the crystal structure of prokaryotic and eukaryotic ribosome-SRP complexes have been solved, the exact mechanism of peptide recognition has not been confirmed, mainly due to the differences in sequence.

The lumen of the ER is an environment full of chaperone proteins which facilitate proper folding before the protein makes its way to the Golgi network from the *cis* face and is subsequently glycosylated and further directed to the cellular membrane with a series of insertion and stop transfer sequences or outside of the cell in a secretory vesicle. These insertion and stop transfer sequences are located between the transmembrane domains and signal the machinery to move to the next portion of transmembrane section of protein or to release the finished synthesis product.

9.2.2 Tail-Anchored Proteins

There is a distinct group of eukaryotic substrates that are delivered to the ER membrane in a post-translational mechanism, called Tail-Anchored proteins (TA). The TA signal is part of the C-terminal domain, so it can only be recognized by cytosolic targeting factors when translation has completed, similar to bacterial T1S. Proteins encoded with the TA signal are destined for the membranes of various organelles and is transported to the respective location by vesicles that bud from the Golgi apparatus.[58] There are two separate ATP-dependent processes that deliver TA proteins to the ER, one uses a heat-shock protein chaperone complex while the other uses a cytosolic ATPase, ASNA1, that also recognizes a signal sequence. Both pathways maintain translocation competence by minimizing misfolds, but the mechanism is relatively unknown beyond the initial substrate recognition event. What is known is that the nature

of the signal sequence seems to be a determinant of the pathway TA proteins take, as more hydrophobic C-terminals favor the ATPase pathway.[5]

9.2.3 Post-translational Modification and Regulation

In many eukaryotes, a simple polypeptide chain is not enough for a protein to be fully functional, and there are a variety of post-translational modifications that have been described in eukaryotic cells. Methylation, acetylation, phosphorylation, and glycosylation are all common modifications that can elicit different recognition or regulation responses. There are countless situations where a single mRNA transcript can form various translation products based on variation in intron splicing or mistranslation from a wobble position in the ribosome, but evidence is now building that identical translation product proteins can be targeted to different parts of the cell based on post-translational modifications,[59] including competition between different modification enzymes for the same modification position.[60]

If a protein is targeted outside of the cell or to the outer membrane, it traverses through the ER to the Golgi network accumulating various modifications and further processing, e.g. the cleavage of proinsulin to form insulin, and the type and extent of modifications varies between species. This process is one of the reasons that prokaryotic systems have trouble producing certain eukaryotic proteins, as they simply lack the appropriate modification pathways. There have been cases where prokaryotes have been transfected to express eukaryotic glycosylation enzymes to produce viable products, but this process can only be done effectively when the modification enzymes are known and non-toxic to the host cell. Considering the premise that excessive genetic manipulation can create a convoluted, possibly nonfunctioning system, a more

reasonable, and promising, method for appropriate post-translational modification in recombinant proteins would be to use eukaryotes themselves.

10 CURRENT SYSTEMS FOR RECOMBINANT PROTEIN SECRETION

Recombinant protein production has been a staple for all types of biotechnology for decades. Industries and universities rely on the availability of specific proteins, but in all the capacity of current molecular biology, the problems of yield and purification are still present. The ability of many organisms to secrete proteins naturally has led to the interest in producing proteins and exporting them outside of the cell, away from large amounts of contaminants and degrading elements and minimizing the physiological impact. The appeal of maintaining a constant state of recombinant protein export without lysing the protein generators is of particular interest to an assortment of industries, such as those manufacturing enzymes for detergents, paper processing, or food production. Medical researchers are interested in the use of cells that can secrete immunogenic proteins and act as live vaccines. Currently, secreted proteins consist of a fraction of the total recombinant protein produced in various areas, but the potential for increased utility is great. This potential can only be realized however with considerable thought and effort into the basic knowledge of secretion systems.

Bacteria are the most efficient producers, but they are unable to perform certain post-translational modifications to eukaryotic cells. This discrepancy in capabilities between secretion systems leads researchers to consider different options when designing a production system. As the target peptide is the factor on which all design plans focus, it's promising to know that protein-based host selection and promoter/

signal peptide optimization are known to yield 3-10 fold increases in specific secretion over wild type strains.[54] In a shift of focus from general cell biology towards a more applicable direction, this chapter outlines a series of model organisms that have been utilized for protein secretion, either for their familiarity, robustness, or secretion capacity.

10.1 *Escherichia coli*

E. coli is a well studied bug, and several reviews have given perspective to its positive and negative qualities as a secretion platform.[9-10,12,61-62] This has made it one of the most popular systems for protein production due to the knowledge of its ability to process a complex number of secretion signals through its variety of export systems.[63] This enthusiasm for *E. coli* was frustrating to an extent, because the species itself does not have an incredible capacity for secretion in general, but directed mutation and the generation of new strains has allowed for novel strategies to be used for an increase in yield by eliminating some of the issues, as is discussed further in Chapter 5. The extent to which *E. coli* has been studied means that its secretion pathways are well characterized, which adds a degree of simplicity to researchers designing a protein production system. Narayanan et al.[14] have compared T1 and T3SSs of *E. coli* in their capacity to export a heterologous lipase from *Pseudozyma antarctica*, named PalB, by fusing the enzyme to HlyA and a flagellar protein gene fliC, and found that the HlyA pathway secreted PalB faster and to a higher extent. It is worth noting that the flagellar T3SS still produced extracellular PalB and there are cases where the flagellar signal sequence gave proteins in higher yield, but is suggested that the similarity to the flagellar protein's sequence is a determining factor.[64] Narayanan also reports that the

extracellular secretion of PalB made disulfide bond formation more effective and the protein less susceptible to proteolysis.

Other secretion signals have been compared in *E. coli*, most as simple fusion products, but others as large constructs. Chung et al. have demonstrated that attaching the T1SS signal lipase ABC transporter recognition domain (LARD) to recombinant proteins can encourage significant amounts of recombinant protein to concentrate extracellularly.[65] The ECUT tag characterized by Paal et al.[66] was designed as a strategy to prevent inclusion body formation and purity of final product. The tag contains a secretion signal from the periplasmic protease inhibitor ecotin fused to a ubiquitin sequence and then the peptide of interest from N to C-terminus. The ubiquitin segment of the fusion product is to allow the fidelity of the target peptide's N-terminus, as deubiquitinating enzymes can be added and cleave the ubiquitin-secretion signal peptide highly selectively away from the peptide of interest.

E. coli also has an advantage over other producers due to its extent of study: a large number of expression systems are known. Some problems with protein production arise when an expression system is not tightly regulated and can lead to unwanted expression at inopportune times, such as different growth phases or has translation rates that are too high or too low. An ideal expression system will only induce the protein of interest when the timing actually calls for it, i.e. only controlling one gene. The expression of a certain protein when it is not appropriate can lead to the formation of inclusion bodies or culture death, and so an expression system without negative side effects is reasonably desired. A few notable expression systems characterized recently contain the ideal qualities of an expression system and have begun to work their way

into a larger scale. The pNEW system characterized by Choi et al.[67] is very tightly regulated and is controlled by a cumate gene switch. The authors claim that, in constrast to the pET system, cells containing this switch remain fully induced for longer time periods, leading to a 2 to 3 fold increase in target protein yields.

10.2 *Streptomyces lividans*

Streptomyces is a second prokaryotic model genus employed in many instances because of its high innate secretion capacity, however their yields for eukaryotic proteins has remained rather low. *S. lividans* is the preferential species host for recombinant protein production because they lack certain impeding processes present in other *Streptomyces* species.[19] There are several proteins critical in biotechnology that are produced at a higher level in *S. lividans* than *E. coli*, namely thermostable phosphotases and cellulases.[68] This is believed to occur because the native signal sequence of the *Thermus* enzymes was able to be processed by *Streptomyces*, and alludes to *S. lividans* being a viable source for many industrially relevant enzymes. A majority of the proteins secreted in *S. lividans* utilize a TAT pathway, which correlates to less bottlenecks at the translocation stage because of the increased concentration of machinery, especially if some naturally secreted proteins are knocked out.[69]

Unfortunately, for several proteins the results for *S. lividans* were low, and the bottlenecks and checkpoints involved remain unidentified. Some ideas regarding the bottlenecks come from the fact that *Streptomyces* has a fairly biased codon usage pattern, due to its high genomic G-C content (roughly 70%),[70] which can clog translation machinery in the search for rare tRNA molecules. In fact, regulation of some *Streptomyces* genes are thought to be influenced by these rare codons, intentionally

slowing the translation of certain genes involved in colony morphogenesis or the metabolism of secondary compounds.[71] Therefore, care must be used if attempting to exchange certain codons or to overexpress the respective tRNAs.

Another challenge in *Streptomyces* protein production is the knowledge of appropriate promoter systems. Many constitutive promoters have been identified, but suitable inducible promoters are still rare. Rodriguez-Garcia et al. have reported the construction of a highly inducible and tightly regulated promoter system for *Streptomyces* using tetracyclin[72] and an ε-caprolactam induced promoter system for a nitrilase gene of *Rhodococcus rhodochrous* has shown promise for use in *Streptomyces*,[73] but further study into these processes is necessary before *Streptomyces* can be used as a more universal protein production system.

10.3 Saccharomyces cerevisiae

When assessing systems for secretion potential, one must inescapably consider fungi. The entire existence of fungi can be explained by their ability to decompose organic matter, and they do this by secreting fairly large amounts of protein. However, the more complicated the organism, the more complicated its secretion pathways, leading most research to be based on yeast and small filamentous fungi.[74] The regulatory systems guiding transcription and translation within eukaryotes tend to limit yields as well as prevent certain heterologous proteins from being expressed.[75] However, these are systems that are incredibly valuable for their ability to post-translationally modify proteins (glycosylation, proper disulfide bond formation, proteolytic editing) and the increasing amount of knowledge regarding these systems, most notably *S. cerevisiae.*[6] The addition of *S. cerevisiae* N-terminal secretory signals to recombinant

proteins has been well established,[76] and when the full pre- and pro- cleavage sites of the signal are intact at translation, they are processed in a stepwise manner during ER and Golgi translocation. On the other hand, *S. cerevisiae* has one of the highest known glycosylation capacities, and can sometimes lead to hyperglycosylated proteins that are targeted for degradation.[77]

In the past few years, many trials of optimization have led to several published (but more commonly proprietary) protocols for protein production by fungi. The most common hurdles in procedural design involve vector and induction choices, as well as regulation of ER stress responses and proteolytic degradation.[54] There are studies that show the half life of most recombinant gene-containing plasmids used by scientists trying to induce protein expression is on the order of hours, and the most effective way to have long-term production of the protein of interest is to actually ligate the gene into the genome of the organism. Yeast are the most effective long term eukaryotic producers of proteins because they are transfected easier than other eukaryotes.

Choice of inducer is a critical point in design because often there is too much of an influx of protein within the ER when transcription and translation of the recombinant protein is switched on and so the folding and transporting machinery inside the ER is overworked and cannot properly fold all of the influxing peptides. This increase in the concentration of misfolded proteins leads to ER stress responses built into the systems. One type of stress response is the unfolded protein response (UPR) that induces the transcription of genes that code for proteins involved in protein folding, modification, and transport. This is not an immediate fix however, and when the ER is under prolonged or excessive stress, it will activate the ER-associated protein degradation (ERAD)

response, which funnels unfolded or misfolded peptides into vacuoles that become lysosomes to digest the excess peptides. This is not an *S. cerevisiae* specific problem, as almost all eukaryotes exhibit some sort of ERAD response, but the high secretion capacity of yeasts is coupled with an effective ERAD system. Another reason for lower yields in fungal production is the fact that most fungi secrete proteases regularly, and these proteases begin to digest recombinant proteins once they both make it out of the cell.

10.4 Pichia pastoris

While *S. cerevisiae* was the first eukaryote used in industrial protein production, *Pichia pastoris* is now the most frequently used yeast species for heterologous protein production.[6] This is because it has a higher secretion capacity as well as a translation efficiency that allows a *P. pastoris* strain with one or two gene copies produce the same amount of extracellular protein as *S.cerevisiae* with 50 copies.[78] This fact especially alludes to further utilization of these yeast for recombinant protein production.

Systems biology has already been used in transcriptomic studies of *P. pastoris* cells expressing recombinant human trypsinogen versus non-induced cells and discovered many secretion helper genes.[79] This in addition to the genome sequencing and analysis which has been done recently[80] leads *P. pastoris* to being a big player in future process development. Temperature studies have also been conducted, and the number of industrial products using *P. pastoris* sourced enzymes is increasing. One downside to its use is its failure to meet food grade requirements, which limits the industrial uses only slightly.

10.5 Aspergillus niger

A filamentous fungus, such as the *Aspergillus* genus, can grow almost indefinitely if there is a steady stream of vesicles moving towards the growing cell tip, or hypha. Filamentous fungi possess the highest known secretion capacity,[81] but if an organism is restricted in its growth, then it will subside from its secretion of protein at the hyphae. To outwit the limiting factors of fungal production, one must be familiar with the big picture. It has been demonstrated that adding amounts of silica or aluminum oxide particles and varying culture conditions can lead to increased branching within *A. niger* batches.[82] Unfortunately at this time, even though their secretion of homologous proteins is unsurpassed by any other system, the production of recombinant proteins has not been able to match those numbers.[6] The hypothesis proposed to explain this phenomenon is the presence of restrictive processing machinery involved in the correct folding and transport of fungal proteins.

Even though *Aspergilli* have been industrially useful for decades, there is still little known about their physiological and genetic characteristics. There has also been a lot of work regarding the regulation or expression of genes that can affect protein folding or transport, and strains have been cultivated that express high levels of chaperone proteins, glycosyltransferases, and export machinery while down regulating the transcription of ER-specific proteases. However, there are enough studies[83] to show that an increased chaperone level or a decreased protease level does not invariably lead to an increase in product recovered.

In addition, *Aspergilli* have been used in food and beverage processes for over 1500 years, in the production of koji foods and cheeses, as well as being a key

producer of enzymes such as glucoamylase for corn syrup refinement and α-galactosidase for digestive supplements. The FDA and WHO have given *Aspergillus nig*er fermentation a generally regarded as safe designation, and this acceptance in the food industry could easily be carried over into the production of medically relevant proteins, especially since they give glycosylation patters similar to mammals, with much lower undesired hyperglycosylation levels compared to other fungal systems.[84]

The downside to *Aspergilli* as a recombinant production platform is the difficulty of transfection. No natural plasmids have proven effective, but stable integration into chromosomal DNA has worked in a way. Transformed DNA fragments are randomly inserted into the chromosome, and plasmid integrations usually results in tandem repeat integrations.[85] The outcome is not easily reproduced or compared. Another complication arises, because gene regulation in fungi occurs at the transcription level, where the gene copy number is essentially independent of expression rate, due to limiting concentrations of transcription factors. However, when one can viably insert inducible promoter sequences (usually carbohydrate based and combined with minimal media growth conditions) into *Aspergilli,* they work effectively for intracellular expression, and are capable of separating growth and production stages.[81] This is not an ideal situation for a secretion system, but the genes of certain transcription factors or translation elements could be included to upregulate appropriate machinery for increased secretion potential. Many promoter systems have been proposed and are being explored, and some of the goals for an ideal promoter would be easily controllable and highly sensitive, such as a metal ion responsive regulator system.[86]

10.6 Insect and Mammalian Platforms

Cell lines of higher organisms, such as silk worms (*Bombyx mori)*, have shown to be successful secretion platforms. Silk worm cells have been successfully infected with a benign baculovirus expression system encoding human interleukins and a silk worm secretion signal sequence. These silk worm cells are of interest because they can form disulfide bonds and when the expression vector is successfully designed, the secretion level can match the expression level. The yields are considerably lower in this system, but it has the ability to synthesize human proteins with minimal discrepancies in post-translational modifications. The signal seems to be dependent on positively charged residues at the N-terminus, and is called signal peptide 1 (SP1) because it is found abundantly secreted into the silkworm haemolymph.[87]

Chinese Hamster Ovary (CHO) cells have also been used recently in the production of high value therapeutic proteins such as monoclonal antibodies and other immune factors. The capability of CHOs to produce human proteins with high homology is undercut by the low secretion levels current constructs have demonstrated. Recombinant protein production levels have increased 100-fold over the past twenty years due to genetic manipulation of chaperone pathways and process control, to the extent that some systems yield g/L scale.[88-89] The point of using these systems is that they can produce pharmaceutical grade protein products, and the only way to achieve this is to direct the products to the secretion pathway. Regrettably, the secretion production levels are currently lower than the intracellular production levels, which do not always lead to fully mature proteins. Work is being done however, and several chaperone proteins have been identified and overexpressed to the effect of increased

secretion and immunogenicity of human immunoglobin G proteins compared to control induced cells.[90]

11 CHALLENGES OF RECOMBINANT PROTEIN PRODUCTION

Protein production has made significant progress over the past decades, but there are still problems that plague both intra- and extracellular production processes. According to Ni et al., there are four main strategies employed for engineering extracellular proteins: Optimizing well characterized secretion systems and fusing the signal to the protein of interest, fusing the protein of interest to a carrier protein without a known translocation mechanism, mutation of certain cell envelope proteins to create leaky cells, or coexpression of a lysis promoting protein.[10] This simplified outlook, while relevant for preliminary troubleshooting, does not fully address some of the more complicated issues that occur between translation and export nor avoids unnecessary cell lysis. This chapter assesses some of the more common issues encountered, with a focus on solutions for extracellular secretion systems.

11.1 Protein Misfolding

A majority of the secretion systems mentioned in Chapters 2 and 3 require an unfolded protein to transverse the secretion channels. T1SSs, by the nature of their C-terminal secretion signal, must export the protein after translation, and in the case of many recombinant proteins without properly expressed chaperone proteins the peptide begins folding upon release from the ribosome, potentially clogging the T1SS transmembrane pore and resulting in secretion machinery bottlenecks.[8] T3SS

dependent proteins, such as flagellar precursors, are also secreted in an unfolded form and fold upon polymerization.[14]

One strategy used in bacterial and yeast systems to improve protein folding is to overexpress different chaperone proteins. This takes an effort to determine which chaperones will be effective, however. BiP, a commonly expressed chaperone in *S. cerevisiae* has been shown to increase the secretion of some proteins while decreasing the secretion of others.[54]

An inherent problem with eukaryotic protein secretion is the complicated mechanism eukaryotic proteins go through before exiting the cell, and each of the steps is regulated and proofread. The UPR and ERAD are more involved as organisms get more complex, and if the flux of protein into the ER is too high, then both are activated. An ideal situation would be to activate the UPR to assist in the proper folding and transport of the newly synthesized peptides without initiating the ERAD, risking the loss of product as the misfolded peptides are degraded. It is promising, however, to note that there have been studies that show that overexpression of UPR chaperones can lead to an increase in protein secretion.[91-92]

11.2 Disulfide Bond Formation

Closely related to the misfolding problem is the incorrect formation of disulfide bonds. If a protein has an erroneous disulfide bond network, it cannot possible fold correctly, but this is separate in consideration because of the fact that a protein can be folded appropriately yet not form the disulfide bonds due to the reducing character of its environment. In prokaryotic systems, the cytoplasm is a reducing environment, and so most proteins that require disulfide bond formation are directed to the periplasm, where

the cysteines can be oxidized.[93] These properties often lead to the misfolding of recombinant proteins that remain in the cytoplasm of the production system. *E. coli* characterized by a mutation in the thioredoxinreductase gene *trxB* were shown to have an inhibitory effect on protein secretion via the hemolysin pathway due to the cytoplasm losing its reductive potential and causing premature disulfide bond formation.[94] However there have been reports of a strain of *E. coli*, named Origami B (DE3) which contains an misexpression of certain cytoplasmic redox enzymes that leads to an oxidizing cytoplasm and fully capable of forming disulfide bonds within the cytoplasm.[95] Strategies for the appropriate formation of disulfide bonds in *E. coli* recombinant proteins have been reviewed extensively by de Marco.[20] Eukaryotic systems have many inherent PDIs in the ER that assure proper disulfide bonds form, and there is evidence that changing the expression levels of certain ones can lead to increase secretion of heterologous proteins.[96]

11.3 Codon Usage and Discrepancies

It has been claimed that codon usage is the single most pressing issue in prokaryotic expression, especially in the attempt to express eukaryotic genes.[70] The *E. coli* strain B834pRareLysS has been demonstrated to provide a higher number of rare tRNAs to the pool within a particular cell to aid in the expression of eukaryotic genes in a prokaryotic system.[97] The Codon Usage Database[98] is an incredible tool to compare the differences in codon frequencies between over 35,000 species, and should be consulted when trying to overcome codon usage issues.

11.4 Other Machinery Bottlenecks

The definition of overexpression states that there is an abundance of a particular protein within a cell, and it follows that if you overexpress a protein that is directed out of the cell without increasing the number of translocation complexes, the system can become saturated. This is easily related to a busy grocery store with only a single open check-out lane. When designing a system for protein expression, one decides on an appropriate signal processing pathway. The post-translational pathway seems to be the most common in yeast protein secretion, but these proteins require a certain amount of chaperone proteins to maintain secretion competence, or a largely unfolded state to allow passage through export channels.[54] Heterologous protein production between prokaryotes and eukaryotes has been difficult because of the difference in or lack of the chaperones. In addition, there are several other limiting factors that can diminish extracellular protein content, such as regulation of transcription and translation, deficiencies in transport machinery, inherent degradation and other quality control mechanisms. There have been several strategies used to minimize certain hang-ups in the protein production process.

In the case of the TAT pathway, the signal sequence must be cleaved before the exported protein can be released from the translocation apparatus. Four different proteases have been identified in *Streptomyces* that recognize the signal sequence, and they are hypothesized to compete for the binding of the precursor proteins and cleave with different efficiencies in a regulatory manner.[19] This could be a lead to pursue when using TAT dependent secretion, as the deletion of a certain signal peptidase that

binds tightly while cleaving inefficiently and the overexpression of a faster cleaving peptidase could greatly affect the translocation rate.

It was mentioned in Section 8.2.2 that the TAT pathway has a low level of protein secretion in normal conditions, but the overexpression of TAT translocation elements can increase the export of GFP,[22] while the deletion or inactivation of the same translocation elements can actually increase secretion via the Sec pathway for some recombinant proteins.[69] The TAT pathway carries peptides across the inner membrane by using the proton-motive force, and it has been shown that the increase of PMF maintenance proteins can increase the secretion of TAT dependent peptides in *Streptomyces*.[99]

Another way to get around the low natural secretion level of the TAT pathway has been to combine a signal fusion sequence with some crafty genetic engineering to produce reasonable amounts of extracellular protein. A deletion of a certain lipoprotein in *E. coli* has given an increase in extracellular protein production, by means of a 'leaky membrane' that allows the recombinant protein to pass through the outer membrane upon direction to the periplasm.[100] There are some concerns to this technique however, as the permeabilization of the outer membrane can seriously deteriorate the physiological state of the cells producing the protein.[14] It is interesting to note, however, that this mutation can cause different effects on cell growth depending on what protein is being expressed, some can seriously diminish growth while others have no significant effect.[100] Various other proteins that can solubilize or permeabilize the outer membrane when overexpressed include Kil and Tol-A in *E. coli*.[20] Others claim that it is easier to simply add a nonionic detergent to the culture medium to lyse the outer membrane upon

direction of protein to the periplasm, but this can cause a decrease in overall production because of the physiological effects of outer membrane lysis diminishes growth.[101]

Many microbes use extracellular proteases for digestion of sustenance, construction of biofilms, or protection from foreign threats. However these proteases can wreck havoc on yields, and so significant effort has been involved in minimizing this factor. The initial strategy was to supplement growth media with protease inhibitors, but this technique can inhibit necessary proteases that are within the cell, leading to an aggregation of dysfunctional proteins within the cell. More recently, genetic manipulation has lead to the develop multi-protease deficient strains of yeast that have up to a 30-fold increase in human growth hormone (hGH) secretion, impressive considering the high proteolytic sensitivity of hGH, while the addition of PDIs and the deletion vacuolar sorting receptor genes in combination with protease deficiency leads to another 50-100% increase in hGH production.[102] Certain *Aspergillus* strains with multiple protease deletions have also resulted in significant increases in recombinant protein secretion.[81]

11.5 Scale-Up

To add to the list of problems one can encounter when designing a protein secretion system, many protocols that produce large amounts of proteins in a shake-flask or other microreactor begin to lose their efficiency when the scale of production increases. Nutrient delivery begins to fall subject to diffusion and uptake rates when the culture volume and populations are larger. Homogenous conditions are difficult to obtain within a reactor, and the physiological attributes of different species can have different effects on bioreactor growth. Many types of bacteria form biofilms at certain cell densities, and these structures can cause havoc on transport processes. Some

secretion systems, such as *Streptomyces* or filamentous fungi species, naturally live a mycelial lifestyle, forming pellets at certain levels of cell density which do not allow proper oxygen and nutrient flow to the center of the mycelium. An approach to this problem is to knock-out expression of genes responsible for peptidoglycan maintenance to discourage correct mycelia formation. This approach has been shown to increase growth rate of *Streptomyces* by 40% and secretion of target proteins 2.5 times wild-type levels.[103] In filamentous fungi, as stated in Section 10.5, have shown to be susceptible to morphology alteration. The addition of talc microparticles increases the formation of hyphal tips, the site of secretion in these species, and thus the secretion capacity. This technique has shown to be effective in both small, shake-flask settings and even more so in larger fed-batch reactors, as Driouch et al. have reported a tenfold improvement in the extracellular production of a fructofuranosidase from *Aspergillus niger*.[104]

Any bioreactor design is going to have to address gas transfer, nutrient mixing, and shear stress.[105] Stirred-tank reactors use an internal propeller system to mix nutrients. Wave bioreactors rock back and forth to mix the contents. Countless other designs are possible, but most use a system for nutrient delivery called a fed-batch. Fed-batch reactors are devices that are well established in biotechnology settings, and are characterized by the carefully monitored administration of nutrients to the culture in an effort to control growth rates and reaction conditions.[106-107] Tightly regulated, inducible promoter sequences are used in typical fed batch reactors, and when cell density has reached the appropriate level, the genes are induced. Fed-batch reactors are ideal systems for inducible promoter sequences because of this inherent ability to control nutrient levels.

Systems biology could be effective in determining factors related to cell growth and division that would need to be controlled in a reactor setting. There is a premise involved with intracellular recombinant protein expression that encourages cell culture growth to a level where cell density is at a maximum before lysis and purification steps take place. However, this premise is counter intuitive to the secretion strategy. The culture, upon induction, exhausts itself synthesizing as much product as possible. When the system is fatigued, the cultures are removed and lysed, the protein of interest is purified, and the system is restarted with fresh media and seed cultures. This would make the secretion argument invalid, because it hinges on the idea that protein extraction is a one-time process. From an export viewpoint, the optimal cell density would be lower than the maximum cell density because there is a need for proper metabolite concentrations if a cell is to remain viable. For continuous protein production and export, the culture medium must not be exhausted of metabolites, and so growth inhibitors will probably need to be used to maintain an appropriate cell density during production stage. This 'appropriate' cell density is most-likely target protein and expression system dependent and further optimization is needed in many instances. In a fed-batch reactor however, the concentration of inhibitors could be easily controlled within the nutrient cycle of the reactor, and used to maintain cell density within a secretion reactor within the effective range without dilution of resources.

One way to deal with particularly fragile or reactive enzymes in a reactor setting is to use a type of scaffolding upon which the enzymes can be linked and immobilized. These enzyme-immobilized membrane bioreactors (EMBRs) are generating interest because they operate by catalyzing a reaction and separating the end product via

solvent flux through the permeable membrane.[108] This strategy can maintain reactivity in some cases, usually reserved for non-aqueous settings without living cultures, but one can imagine the construction of a multi-chambered EMBR that has been designed to build metabolites or pharmaceutical compounds in a step-wise process through various enzyme-linked membranes. This type of system is not proposed for a secretion-based production reactor but a post-purification application for the proteins that are produced in a secretion system.

12 PRACTICAL APPLICATIONS

There are plenty of reasons to pursue protein export as a means to large scale production. In any economic climate, purchasing power increases on a macro level when the prices of certain commodities drop. In the current political climate, the converse of this premise is being applied to arguments regarding the correlation between rising health and energy costs and the abjection of our national economy and security. Recombinant proteins are being used more and more as alternative therapeutics, materials, and energy production mechanisms, but at this time these technologies can not compete economically at scale without subsidies or vouchers.

One way to improve production costs is to increase the efficiency of the production process, and, as stated above, using extracellular secretion of recombinant proteins is a promising method for that end. Of course, much more research needs to be done in almost all of the areas which could employ secretion pathways at the production level, but this chapter is a display of instances where secretion of certain proteins has demonstrated a proof of concept of economic and/or societal benefit.

12.1 Pharmaceutical Production

The human immune system, to briefly overview, can be described by two mechanisms: The innate response and the adaptive response. The innate response is non-specific and protects the body against pathogens. It is responsible for initiating the adaptive response by recruiting immune cells to the point of infection so that an adaptive, specific response can begin. The adaptive response occurs when antigens (small, pathogen specific peptides) are presented to T-cells by antigen presenting cells (APCs) or dendritic cells (DCs; APCs that also express immune system stimulating factors) which triggers a signal cascade ending in the clonal expansion of the antigen specific, activated T-cell. This army of T-cell clones attacks the foreign entity and then, upon victory, downsize the population to a select few memory T-cells that patrol the body searching for a reappearance of the specific antigen it recognizes.[109] Modern therapeutics are looking for ways to use these processes in conjunction with pharmaceutical compounds to minimize side effects and increase the effectiveness of certain treatments.

One strategy has been to create soluble T-cell receptors, as opposed to membrane bound and disulfide-linked structures, connected by a flexible segment to a specific antigen peptide that can bind to the receptor's active site, because the natural form has not been successfully expressed and can reversibly bind to the antigen.[97] This approach works well for x-ray crystallography studies, but *in vivo* the T-cells must also bind to the major histocompatibility complex (MHC) of APCs and DCs for activation to occur. Soluble MHCs, however, are becoming more popular targets for recombinant production,[110] and the hypothesis states that soluble MHC: antigen complexes could

travel through the blood and lymph systems priming T-cells for activation against the antigen. If structural studies can confirm homology and immunosorbent competence, secretion pathways could definitely be employed for pharmaceutical grade production of these proteins for therapeutic trials. Antibodies are essentially soluble B-cell receptors that follow a similar activation pathway as T-cells, and they balance out the immune system's responsiveness and specificity. Fernandez et al. demonstrated how both short-chain variable fragments (scFvs) and heavy-chain variable fragments (V_HH) can be secreted as fusion constructs with hemolysin C-terminal sequences in their fully oxidized, functional form.[94,111]Another method proven to produce functional, fully glycosylated antibodies uses an engineered signal sequence based on the α-mating factor (αMF1) of *S*. *cerevisiae*, when combined with strain optimization, led to significant a 180-fold increase in extracellular production of human IgG1.[112] The αMF1 is a fairly conserved secretion signal in eukaryotes and consists of two cleavage points; one for the ER, one for the Golgi. It has been used a fair amount for the secretion of recombinant proteins, but its success as a secretion signal greatly depends on its fusion partner.[113]

One issue is that scFvs contain at least one disulfide bond in each chain, and this must be considered when producing these molecules. A creative way around this problem and to allow scFvs to be produced more effectively in prokaryotic systems is to mutate the protein to fold without the disulfide bonds, which has been implemented in recombinant protein production but not yet demonstrated via a secretion pathway.[114-115]

Hormones are another type of biomolecules that can be used as therapies. They are small peptides that initiate various responses within the body. One of the more

highly produced hormones is insulin, which is reasonable because as diabetes cases continue to rise, the demand for insulin and proinsulin have increased dramatically. *In vivo* production of insulin proceeds through several steps of folding and cleavage from an initial translation product into the heterologous dimeric active form. In current industrial processes, proinsulin has been fused to the *E. coli* protein ecotin and shown to form correctly in the periplasm of *E. coli*, but the naturally occurring proteases within the periplasm still keep yields low in *E. coli*.[116] In *S. cerevisiae* proinsulin production is hampered by hyperglycosylation and inability of the yeast to successfully process the human proinsulin.[117] However, it has been reported that proinsulin-like peptides without certain glycoslyation sites can be processed by the secretion pathway and an *in vitro* cleavage step will yield a semi-functional insulin molecule.[118] The use of microbes to produce insulin is gaining support from the markets as well, considering that most pharmaceutical insulin is isolated from pigs, and religious beliefs cause some diabetes sufferers to deny porcine-sourced medicines. Additionally, the demand for insulin is exceeding what the farming industry can provide.

Other chemokines responsible for directing cell transport within the body, such as inflammatory responses or angiogenesis, have been the target for drug discovery in recent years, and development has been made into their production within prokaryotic systems. The main challenge with their production is maintaining an intact N-terminal sequence after purification, and a few strategies have worked well in intracellular expression.[95] The inclusion of a ubiquitin tag within the fusion product, as described by Paal et al.,[66] could be used in secretion and then accurate cleavage of the cytokines away from the fusion construct.

12.2 Live-Vaccine Therapeutics

The bulk production of pharmaceuticals via recombinant protein export is an extremely promising concept which is in various stages of scaling up, depending on what product is being discussed. However, there are more delicate medical problems that could benefit from protein export and a different approach. The T3 and T4SSs is a very complicated process with tight regulation of timing and structure development,[119] which would be inefficient as a bulk protein production pathway, but cancer, hemophilia, sickle-cell anemia, and afflictions based on other genetic misregulations could be targeted by pathogen-like bacteria expressing antigens or inhibitory peptides that can be directed into infected cells.

The use of bacteria in cancer therapy has taken two different routes. First, bacteria are systematically or directly administered to the tumor itself, and bacterial replication within tumor cells leads to tumor regression. The second uses bacteria as antigen delivery particles to develop cancer specific immune responses. Some strains of *Salmonella* offer a potential means to deliver antigens to $CD8^+$ T-cells to stimulate cytotoxic activity against tumor cells. This system has even been claimed to sensitize tumor cells to preexisting, circulating $CD8^+$ cells.[120] This technique has been used in both cancers with and without a known infectious origin, and if a proper peptide is administered with a high enough efficiency it can theoretically induce an immune response and would be an important therapeutic strategy because it would not require full genetic knowledge of the tumor's antigenic content.[120] Avirulent bacteria vectors, pathogens with certain virulence genes deleted, offer a reasonable approach for this type of therapy. Their genetics are understood and relatively easy to manipulate, to the

extent that the strains can be engineered to express several antigens. They can also be grown and administered cheaply, and can act as their own adjuvant to initiate an immune response.[121] Specifically, avirulent *Salmonella* has recently been used to target cancer cells and other autoimmune disorders through its T3SS and the expression of antigens or therapeutics. The extent of study so far has been reviewed by Moreno et al.[122]

Recently, a drive towards the use of T5SSs to display certain recombinant proteins in the outer membrane of some gram-negative bacteria has shown to work rather effectively, much more effectively than the use of T5SS for recombinant protein production.[124] This could be another potential route to design a live-vaccine that displays certain human cell-surface proteins in an attempt to prolong the *in vivo* half-life of these cultures by immune system evasion. Obviously careful design strategies must be employed to discourage the proliferation of virulent, immune-evading pathogens, but the allure of more effective therapies warrants further investigation. In addition, since the T4SS has the capability to transport peptides, protein-protein complexes, and nucleic acid-protein complexes,[123] one can imagine the creation of highly specialized live-vaccine cells using two different secretion pathways for different products to minimize competition for the transport machinery.

12.3 Energy Production

Cellulose is the most abundant biopolymer on the planet, and thus being so, it currently has several applications as a renewable resource in many industries, including paper and clothing production, the stationary phase in most thin-layer chromatography plates, and “green” building insulation.[125] Its sheer abundance and the relative ease of

replacement gives cellulose enormous potential as an energy source. The chemical composition of cellulose is relatively simple, as the polymer consists of several thousand glucose residues bound together in β-1,4-glycosidic linkages forming long strands that are capable of forming hydrogen bonds with neighboring strands that increase its stability as well as prohibiting its solubility in water. Glucose is a very energy-rich compound, and this energy can be used by industries via production of biofuels or other organic compounds. The most talked about product is ethanol, formed during glucose fermentation and a widely used gasoline additive. Other alcohols, such as butanol, with a higher energy density have the potential to be a replacement altogether. The ability to produce ethanol from cellulosic waste in a cheap and efficient manner would be of great importance. There are three types of cellulolytic enzymes that are commonly found in nature, and all three are necessary to efficiently break cellulose down into its glucose monomers. Every enzyme that hydrolyses cellulose does so by cleaving the β-1,4-glycosidic bond between glucose subunits, but the substrate specificity is the distinguishing factor. The enzyme β-glucanase, member of the endoglucanase family, cleaves internal glycosidic bonds and breaks interstrand hydrogen bonds, thereby enhancing the solubility of cellulose and cleaving strands of random length from ten to a few hundred glucose subunits. The cellooligohydrolase family that cleaves branches sized from two to ten glucose subunits from the reducing ends of cellulose. The last type are called β-glucosidases, of the exoglucanase family of enzymes, and they cleave glucose monomers from the reducing ends of cellulose fibers.[126]

The export of a mixture of these enzymes is normal in the everyday metabolism of certain bacteria and fungi, and the engineering of hypersecretion strains with the development of efficient cellulase ratios could bring glucose production from cellulose feed stocks into a more economically competitive process. Various proteins have been identified that enhance recombinant glucanase activity,[127] and genetic manipulation of these proteins as well as secretory machinery or helper proteins could presumably serve as a starting point in a cellulolytic system design. Once the glucose units are produced from the cellulose, there are countless products that could be produced with all sorts of metabolic pathways.

Alcohol fuels are easier to produce, and the level of corn and other feedstock subsidies make alcohols very tempting to pursue, but they do not have as high an energy density as long chain alkanes such as those used in jet or diesel fuel. Recently, bacteria have been exhibited to produce a wide variety of fatty acids in response to various environmental stressors; the ratio of saturated to unsaturated can vary based on growth temperature or the presence of ethanol or other fatty acids in the growth medium.[128] These fatty acids can be converted to petroleum-like chemicals like long chain alkanes for fuel or other industrial purposes. There are a few strategies being employed that use different metabolic pathways, which have been usefully reviewed by Yan et al.[129] Most recently, however, the typical strategy for producing these long chain alkanes has consisted of engineering the fatty-acid biosynthesis pathway to overexpress the subunits of acetyl-CoA carboxylase to increase the rate of malonyl-CoA synthesis, which is widely accepted as the rate-limiting step of fatty acid biosynthesis, in combination with the deletion of several β-oxidation enzymes that are responsible for

degrading fatty acids if their concentration gets too high. Upon the completed synthesis of a C_{12} or C_{14} fatty acid, esterification and decarboxylation are the two most common ways to obtain a viable fuel product. However, both of these products require the lysis of the cultured cells and the extraction of the fatty acids before catalytically creating the end product.[130] The other approach is to use the cultured cells to complete the conversion as the last step in the process. Certain alkane production pathways have been discovered in cyanobacteria that have the capability to synthesize C_{13} to C_{17} alkane mixtures and secrete them out of the cell. The main enzyme is an aldehyde decarbonylase which removes a carbon monoxide molecule from the end of a long chain fatty aldehyde previously reduced from a fatty acid. This pathway has been heterologously expressed in *E. coli* to the extent that alkane levels were comparable to the wild-type cyanobacteria.[131] The difference in production organism centers on the source of nutrition. *E. coli* needs some sort of carbon feedstock to produce the alkanes via respiration while the cyanobacteria, such as *Anabaena* or *Synechoccus* species, use CO_2 and light in photosynthesis. Both processes have scalability issues: E. coli requires a certain oxygen level and continuous nutrient flow while high cyanobacteria cell density greatly diminishes the transmittance of light into the reactor and can affect CO_2 uptake rates as well.[132]

There is significant debate as to which method will prove to be the most efficient and scalable. Some say that the production of alkanes by the microbes themselves will lead to lower yields because of increased metabolic demands, while others argue that the loss of product during the conversion processes are comparable to the lower secretion yields. In reality, none of these methods is close to replacing petroleum based

long chain alkanes as a fuel source. The processes are just being described in academically useful ways, and to date no one has published a report of extracellular alkane synthesis using membrane bound or extracellular proteins. It is exciting to think of this frontier being explored in the coming years now that heterologous recombinant alkane biosynthetic enzymes have been achieved.

12.4 Spider Silk Monomers

Spider silk is one of nature's most spectacular materials. Stronger than steel by weight, they have a large amount of potential uses including medical implants, high-strength fibers, and drug delivery systems.[133] However, the production of native silk proteins in a recombinant manner has been problematical through several factors. First, the genes of silk proteins contain large amounts of repeat regions, which lead to homologous recombination. Second, silk gene codon usage is specialized within the silk producing cells of spiders, which can lead to translation stalling and incomplete protein production when recombinant expression is employed. Third, a high concentration of these proteins within a cell can cause the monomers to polymerize, forming fibers within the cell and destroying its ability to continue producing the protein. Widmaier et al.[134] have developed an optimized procedure for minimizing these factors with a system that consists of four genetic parts: a signal sequence, an affinity tag, a cleavage signal, and the spider silk protein. The construct is designed for *E. coli* and it uses a salt-dependent inducer with low basal transcription rate and no other activity under normal conditions. It is also set up to restrict translation of the recombinant protein until after the T3SS translocation apparatus has been constructed to negate the formation of inclusion bodies.

The idea of using spider silk as a material for everyday use is promising, and in an ideal situation the monomers could be secreted in their native form to produce a mixture that could be 'spun' in an industrially favorable way. This is a little bit further out than the other practical applications mentioned in this chapter, but the research is happening.

13 POTENTIAL AVENUES FOR FUTURE RESEARCH

The process of designing an expression system is inherently dependent on the target. The purity and amounts of the said target needed are also requirements in the design process, dependent on the end application of the recombinant protein as pharmaceutical grade products must be more highly purified than other industrial proteins. The synthesis of recombinant proteins can be an arduous task, as the diversity of proteins inevitably leads to protocols that either work with minimal tweaking or procedures that need months of optimization. Past research has revealed the daunting task that is designing a highly efficient system for the production of a specific protein. Still, technology is advancing at a promising rate. Sequencing techniques allow for characterizations on a cell to cell basis, and systems biology approaches have been used to assess the random mutagenesis of various microbes in an attempt to grow strains with desirable properties and then transfect them with recombinant genes directed to secretion systems.[135-136] There is still, however, a fair number of proteins with industrial relevance that have not yet been expressed within a secretion pathway. The strategies for effective and efficient protein secretion should center on the idea that bacteria should be used to produce bacterial proteins while eukaryotic proteins should

be produced in eukaryotic systems. There are cases where a bacterial system has successfully secreted eukaryotic proteins, but in general the bottlenecks involved with trans-domain protein production drastically diminish the yield and quality of the product. When designing a protein secretion system, one should keep this in mind.

Systems biology has had a large impact on the design and characterization of secretion systems to date, but the potential of these techniques to quantify and analyze the effectiveness of a system has not been actualized, though coding strategies are being investigated.[137] It would be very beneficial to identify the larger network of genes and proteins that are involved in various product- or species-specific secretion mechanisms because strains of microbes could be designed with optimized expression levels of various chaperones, PDIs, ribosomal elements and export machines towards highly efficient processes. Further study into feedback pathways and regulatory elements of secretion platforms is desperately needed to expose the current unknowns that hinder output.

In the process of studying processes on a cell-wide scale, the bioinformatics field has obtained incredibly large amounts of data over the past few years, and the task of digging through multiple databases is daunting. The Wikipedia “List of Biological Databases” has over 150 items, constantly changing as old databases are no longer maintained and new ones take their place, in a process that is full of redundancies.[138-139] The compilation of the acquired data also needs to occur in a highly organized and accessible manner for ease of interpretation and collaboration.

While it is currently impractical to think of buildings and bridges made of spider silk or airplanes that run on microbial fuel cells, and cost-benefit analysis shows that

intracellular expression and the subsequent purification is still more effective for certain proteins than extracellular secretion,[140] advances occur every day that allow for higher production levels. There is a minor consensus that the eventual goal of is to develop a strain of microbe that can continuously secrete large amounts of a wide range of proteins, due to its expression of an appropriate level of chaperones and PDIs, minimal protease activity, an excess of transport machinery, and (if eukaryotic) optimized post-translational modification pathways. This simple and generic production system has been called the ‘Holy Grail’ by some molecular biologists, but when one considers the actual behavior and incredible diversity of secretion mechanisms, a single system that can accommodate such a wide range seems like wishful thinking. On the contrary, the diversity of secretion systems described in nature can be used to the extent that almost any scientist can find a system that will work effectively for their target protein. One just has to look around.

14 REFERENCES

1 Crick, F. Central Dogma of Molecular Biology. *Nature* **227**, 561-563, (1970).

2 Feng, J. L. *et al.* The RNA Component of Human Telomerase. *Science* **269**, 1236-1241, (1995).

3 Fire, A. *et al.* Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* **391**, 806, (1998).

4 Liang, J. & Dill, K. A. Are Proteins Well-Packed? *Biophys. J.* **81**, 751-766, (2001).

5 Cross, B. C. S., Sinning, I., Luirink, J. & High, S. Delivering proteins for export from the cytosol. *Nature Reviews Molecular Cell Biology* **10**, 255-264, (2009).

6 Schmidt, F. R. Recombinant expression systems in the pharmaceutical industry. *Appl. Microbiol. Biotechnol.* **65**, 363-372, (2004).

7 Simonen, M. & Palva, I. Protein secretion in *Bacillus* species. *Microbiol. Mol. Biol. Rev.* **57**, 109-137, (1993).

8 Delepelaire, P. Type I secretion in gram-negative bacteria. *Biochim. Biophys. Acta-Mol. Cell Res.* **1694**, 149-161, (2004).

9 Mergulhão, F. J. M., Summers, D. K. & Monteiro, G. A. Recombinant protein secretion in *Escherichia coli*. *Biotechnol. Adv.* **23**, 177-202, (2005).

10 Ni, Y. & Chen, R. Extracellular recombinant protein production from *Escherichia coli*. *Biotechnol. Lett* **31**, 1661-1670, (2009).

11 Binet, R., Letoffe, S., Ghigo, J. M., Delepelaire, P. & Wandersman, C. Protein secretion by Gram-negative bacterial ABC exporters - A review. *Gene* **192**, 7-11, (1997).

12 Shokri, A., Sanden, A. M. & Larsson, G. Cell and process design for targeting of recombinant protein into the culture medium of *Escherichia coli*. *Appl. Microbiol. Biotechnol.* **60**, 654-664, (2003).

13 Cianciotto, N. P. Type II secretion: a protein secretion system for all seasons. *Trends Microbiol.* **13**, 581-588, (2005).

14 Narayanan, N., Khan, M. & Chou, C. P. Enhancing functional expression of heterologous lipase B in *Escherichia coli* by extracellular secretion. *J. Ind. Microbiol. Biotechnol.* **37**, 349-361, (2010).

15 Wild, J., Altman, E., Yura, T. & Gross, C. A. DnaK and DnaJ Heat-Shock Proteins Participate in Protein Export in *Eschericia coli*. *Genes Dev.* **6**, 1165-1172, (1992).

16 Wild, J., Rossmeissl, P., Walter, W. A. & Gross, C. A. Involvement of the DnaK-DnaJ-GrpE chaperone team in protein secretion in *Escherichia coli*. *J. Bacteriol.* **178**, 3608-3613, (1996).

17 Kaderbhai, N. N., Ahmed, K. & Kaderbhai, M. A. Export of a hyperexpressed mammalian globular cytochrome b(5) precursor in *Escherichia coli* is dramatically affected by the nature of the amino acid flanking the secretory signal sequence cleavage bond. *Protein Sci.* **19**, 1344-1353, (2010).

18 Valent, Q. A. *et al.* The *Escherichia coli* SRP and SecB targeting pathways converge at the translocon. *EMBO J.* **17**, 2504-2512, (1998).

19 Vrancken, K. & Anne, J. Secretory production of recombinant proteins by *Streptomyces*. *Future Microbiol* **4**, 181-188, (2009).

20 de Marco, A. Strategies for successful recombinant expression of disulfide bond-dependent proteins in *Escherichia coli*. *Microbial Cell Factories* **8**, (2009).
21 Xia, Y. *et al.* Extracellular secretion in *Bacillus subtilis* of a cytoplasmic thermostable beta-galactosidase from *Geobacillus stearothermophilus*. *J. Dairy Sci.* **93**, 2838-2845, (2010).
22 Barrett, C. M. L., Ray, N., Thomas, J. D., Robinson, C. & Bolhuis, A. Quantitative export of a reporter protein, GFP, by the twin-arginine translocation pathway in *Escherichia coli*. *Biochem. Biophys. Res. Commun.* **304**, 279-284, (2003).
23 Pop, O., Martin, U., Abel, C. & Muller, J. P. The twin-arginine signal peptide of PhoD and the TatA(d)/C-d proteins of Bacillus subtilis form an autonomous tat translocation system. *J. Biol. Chem.* **277**, 3268-3273, (2002).
24 Sandkvist, M. Type II secretion and pathogenesis. *Infect. Immun.* **69**, 3523-3535, (2001).
25 Parche, S., Geissdorfer, W. & Hillen, W. Identification and characterization of xcpR encoding a subunit of the general secretory pathway necessary for dodecane degradation in *Acinetobacter calcoaceticus* ADP1. *J. Bacteriol.* **179**, 4631-4634, (1997).
26 De Groot, A. *et al.* Exchange of Xcp (Gsp) secretion machineries between *Pseudomonas aeruginosa* and *Pseudomonas alcaligenes*: Species specificity unrelated to substrate recognition. *J. Bacteriol.* **183**, 959-967, (2001).
27 de Vrind, J., de Groot, A., Brouwers, G. J., Tommassen, J. & de Vrind-de Jong, E. Identification of a novel Gsp-related pathway required for secretion of the manganese-oxidizing factor of *Pseudomonas putida* strain GB-1. *Mol. Microbiol.* **47**, 993-1006, (2003).
28 DiChristina, T. J., Moore, C. M. & Haller, C. A. Dissimilatory Fe(III) and Mn(IV) reduction by *Shewanella putrefaciens* requires *ferE*, a homolog of the *pulE* (*gspE*) type II protein secretion gene. *J. Bacteriol.* **184**, 142-151, (2002).
29 Arrieta, J. G. *et al.* A type II protein secretory pathway required for levansucrase secretion by *Gluconacetobacter diazotrophicus*. *J. Bacteriol.* **186**, 5031-5039, (2004).
30 Deane, J. E., Abrusci, P., Johnson, S. & Lea, S. M. Timing is everything: the regulation of type III secretion. *Cell. Mol. Life Sci.* **67**, 1065-1075, (2010).
31 Anderson, D. M. & Schneewind, O. A mRNA signal for the type III secretion of Yop proteins by *Yersinia enterocolitica*. *Science* **278**, 1140-1143, (1997).
32 Karavolos, M. H., Wilson, M., Henderson, J., Lee, J. J. & Khan, C. M. A. Type III secretion of the *Salmonella* effector protein SopE is mediated via an N-terminal amino acid signal and not an mRNA sequence. *J. Bacteriol.* **187**, 1559-1567, (2005).
33 Dobo, J. *et al.* Application of a Short, Disordered N-Terminal Flagellin Segment, a Fully Functional Flagellar Type III Export Signal, to Expression of Secreted Proteins. *Appl. Environ. Microbiol.* **76**, 891-899, (2010).
34 Lloyd, S. A., Forsberg, A., Wolf-Watz, H. & Francis, M. S. Targeting exported substrates to the *Yersinia* TTSS: different functions for different signals? *Trends Microbiol.* **9**, 367-371, (2001).
35 McDermott, J. E. *et al.* Computational Prediction of Type III and IV Secreted Effectors in Gram-Negative Bacteria. *Infect. Immun.* **79**, 23-32, (2011).

36 Pallen, M. J., Beatson, S. A. & Bailey, C. M. Bioinformatics, genomics and evolution of non-flagellar type-III secretion systems: a Darwinian perpective. *FEMS Microbiol. Rev.* **29**, 201-229, (2005).

37 Negrea, A. *et al.* Salicylidene acylhydrazides that affect type III protein secretion in *Salmonella enterica serovar Typhimurium*. *Antimicrob. Agents Chemother.* **51**, 2867-2876, (2007).

38 Wilharm, G., Dittmann, S., Schmid, A. & Heesemann, J. On the role of specific chaperones, the specific ATPase, and the proton motive force in type III secretion. *Int. J. Med. Microbiol.* **297**, 27-36, (2007).

39 Bates, S., Cashmore, A. M. & Wilkins, B. M. IncP plasmids are unusually effective in mediating conjugation of *Escherichia coli* and *Saccharomyces cerevisiae*: Involvement of the Tra2 mating system. *J. Bacteriol.* **180**, 6538-6543, (1998).

40 Waters, V. L. Conjugation between bacterial and mammalian cells. *Nat. Genet.* **29**, 375-376, (2001).

41 Cascales, E. & Christie, P. J. The versatile bacterial type IV secretion systems. *Nature Reviews Microbiology* **1**, 137-149, (2003).

42 Thanassi, D. G., Stathopoulos, C., Karkal, A. & Li, H. L. Protein secretion in the absence of ATP: the autotransporter, two-partner secretion and chaperone/usher pathways of Gram-negative bacteria (Review). *Mol. Membr. Biol.* **22**, 63-72, (2005).

43 Brandon, L. D. & Goldberg, M. B. Periplasmic transit and disulfide bond formation of the autotransported *Shigella* protein IcsA. *J. Bacteriol.* **183**, 951-958, (2001).

44 Bulieris, P. V., Behrens, S., Holst, O. & Kleinschmidt, J. H. Folding and insertion of the outer membrane protein OmpA is assisted by the chaperone Skp and by lipopolysaccharide. *J. Biol. Chem.* **278**, 9092-9099, (2003).

45 Filloux, A., Hachani, A. & Bleves, S. The bacterial type VI secretion machine: yet another player for protein transport across membranes. *Microbiology-(UK)* **154**, 1570-1583, (2008).

46 Lesic, B., Starkey, M., He, J., Hazan, R. & Rahme, L. G. Quorum sensing differentially regulates *Pseudomonas aeruginosa* type VI secretion locus I and homologous loci II and III, which are required for pathogenesis. *Microbiology-(UK)* **155**, 2845-2855, (2009).

47 Weber, B., Hasic, M., Chen, C., Wai, S. N. & Milton, D. L. Type VI secretion modulates quorum sensing and stress response in *Vibrio anguillarum*. *Environ. Microbiol.* **11**, 3018-3028, (2009).

48 Schwarz, S. *et al. Burkholderia* Type VI Secretion Systems Have Distinct Roles in Eukaryotic and Bacterial Cell Interactions. *PLoS Path.* **6**, (2010).

49 Schwarz, S., Hood, R. D. & Mougous, J. D. What is type VI secretion doing in all those bugs? *Trends Microbiol.* **18**, 531-537, (2010).

50 Abdallah, A. M. *et al.* Type VII secretion - mycobacteria show the way. *Nature Reviews Microbiology* **5**, 883-891, (2007).

51 Simeone, R., Bottai, D. & Brosch, R. ESX/type VII secretion systems and their role in host-pathogen interaction. *Curr. Opin. Microbiol.* **12**, 4-10, (2009).

52 Tseng, T.-T., Tyler, B. & Setubal, J. Protein secretion systems in bacterial-host associations, and their description in the Gene Ontology. *BMC Microbiol.* **9**, S2, (2009).

53 Xu, D. & Esko, J. D. A Golgi-on-a-chip for glycan synthesis. *Nat. Chem. Biol.* **5**, 612-613, (2009).

54 Idiris, A., Tohda, H., Kumagai, H. & Takegawa, K. Engineering of protein secretion in yeast: strategies and impact on protein production. *Appl. Microbiol. Biotechnol.* **86**, 403-417, (2010).

55 Lodish, H. *et al. Molecular Cell Biology*. 4th Ed. edn, (W.H. Freeman and Co., 2000).

56 Shen, H. B., Yang, J. & Chou, K. C. Euk-PLoc: an ensemble classifier for large-scale eukaryotic protein subcellular location prediction. *Amino Acids* **33**, 57-67, (2007).

57 Luirink, J. & Sinning, I. SRP-mediated protein targeting: structure and function revisited. *Biochim. Biophys. Acta-Mol. Cell Res.* **1694**, 17-35, (2004).

58 Borgese, N., Colombo, S. & Pedrazzini, E. The tale of tail-anchored proteins. *The Journal of Cell Biology* **161**, 1013-1019, (2003).

59 Karniely, S. & Pines, O. Single translation-dual destination: mechanisms of dual protein targeting in eukaryotes. *EMBO Rep.* **6**, 420-425, (2005).

60 Ryan, K. & Bauer, D. L. V. Finishing touches: Post-translational modification of protein factors involved in mammalian pre-mRNA 3 ' end formation. *International Journal of Biochemistry & Cell Biology* **40**, 2384-2396, (2008).

61 Gentschev, I. *et al.* A 40 years encounter with *Escherichia coli*. *Nova Acta Leopold.* **98**, 41-45, (2008).

62 Sommer, B. *et al.* Extracellular production and affinity purification of recombinant proteins with *Escherichia coli* using the versatility of the maltose binding protein. *J. Biotechnol.* **140**, 194-202, (2009).

63 Pugsley, A. P., Kornacker, M. G. & Poquet, I. The general protein-export pathway is directly required for extracellular pullulanase secretion in *Escherichia coli* K12. *Mol Microbiol* **5**, 343-352, (1991).

64 Majander, K. *et al.* Extracellular secretion of polypeptides using a modified *Escherichia coli* flagellar secretion apparatus. *Nat. Biotechnol.* **23**, 475-481, (2005).

65 Chung, C. W. *et al.* Export of recombinant proteins in *Escherichia coli* using ABC transporter with an attached lipase ABC transporter recognition domain (LARD). *Microbial Cell Factories* **8**, (2009).

66 Paal, M., Heel, T., Schneider, R. & Auer, B. A novel Ecotin-Ubiquitin-Tag (ECUT) for efficient, soluble peptide production in the periplasm of *Escherichia coli*. *Microbial Cell Factories* **8**, (2009).

67 Choi, Y. J. *et al.* Novel, Versatile, and Tightly Regulated Expression System for *Escherichia coli* Strains. *Appl. Environ. Microbiol.* **76**, 5058-5066, (2010).

68 Diaz, M. *et al.* High-level overproduction of *Thermus* enzymes in *Streptomyces lividans*. *Appl. Microbiol. Biotechnol.* **79**, 1001-1008, (2008).

69 De Keersmaeker, S. *et al.* Functional analysis of TatA and TatB in *Streptomyces lividans*. *Biochem. Biophys. Res. Commun.* **335**, 973-982, (2005).

70 Gustafsson, C., Govindarajan, S. & Minshull, J. Codon bias and heterologous protein expression. *Trends Biotechnol.* **22**, 346-353, (2004).

71 Nguyen, K. T. *et al.* Colonial differentiation in *Streptomyces coelicolor* depends on translation of a specific codon within the adpA gene. *J. Bacteriol.* **185**, 7291-7296, (2003).

72 Rodriguez-Garcia, A., Combes, P., Perez-Redondo, R., Smith, M. C. A. & Smith, M. C. M. Natural and synthetic tetracycline-inducible promoters for use in the antibiotic-producing bacteria *Streptomyces*. *Nucleic Acids Res.* **33**, 8, (2005).

73 Herai, S. *et al.* Hyper-inducible expression system for streptomycetes. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 14031-14035, (2004).

74 Archer, D. B. & Peberdy, J. F. The molecular biology of secreted enzyme production by fungi. *Crit. Rev. Biotechnol.* **17**, 273-306, (1997).

75 Peberdy, J. F. Protein Secretion in Filamentous Fungi - Trying to Understand a Highly Productive Black-Box. *Trends Biotechnol.* **12**, 50-57, (1994).

76 Brake, A. J. *et al.* Alpha-factor-directed synthesis and secretion of mature foreign proteins in *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences* **81**, 4642-4646, (1984).

77 Gellissen, G. & Hollenberg, C. P. Application of yeasts in gene expression studies: A comparison of *Saccharomyces cerevisiae*, *Hansenula polymorpha* and *Kluyveromyces lactis* - a review. *Gene* **190**, 87-97, (1997).

78 Buckholz, R. G. & Gleeson, M. A. G. Yeast Systems for the Commercial Production of Heterlogous Proteins. *Bio-Technology* **9**, 1067-1072, (1991).

79 Gasser, B., Sauer, M., Maurer, M., Stadlmayr, G. & Mattanovich, D. Transcriptomics-based identification of novel factors enhancing heterologous protein secretion in Yeasts. *Appl. Environ. Microbiol.* **73**, 6499-6507, (2007).

80 De Schutter, K. *et al.* Genome sequence of the recombinant protein production host *Pichia pastoris*. *Nat. Biotechnol.* **27**, 561-U104, (2009).

81 Fleissner, A. & Dersch, P. Expression and export: recombinant protein production systems for *Aspergillus*. *Appl. Microbiol. Biotechnol.* **87**, 1255-1270, (2010).

82 Driouch, H., Sommer, B. & Wittmann, C. Morphology engineering of *Aspergillus niger* for improved enzyme production. *Biotechnol. Bioeng.* **105**, 1058-1068, (2010).

83 Sharma, R., Katoch, M., Srivastava, P. S. & Qazi, G. N. Approaches for refining heterologous protein production in filamentous fungi. *World J. Microbiol. Biotechnol.* **25**, 2083-2094, (2009).

84 Maras, M., van Die, I., Contreras, R. & van den Hondel, C. Filamentous fungi as production organisms for glycoproteins of bio-medical interest. *Glycoconjugate J.* **16**, 99-107, (1999).

85 Weld, R. J., Plummer, K. M., Carpenter, M. A. & Ridgway, H. J. Approaches to functional genomics in filamentous fungi. *Cell Res.* **16**, 31-44, (2006).

86 Winge, D. R., Jensen, L. T. & Srinivasan, C. Metal-ion regulation of gene expression in yeast. *Curr. Opin. Chem. Biol.* **2**, 216-221, (1998).

87 Futatsumori-Sugai, M. & Tsumoto, K. Signal peptide design for improving recombinant protein secretion in the baculovirus expression vector system. *Biochem. Biophys. Res. Commun.* **391**, 931-935, (2010).

88 Lim, Y. *et al.* Engineering mammalian cells in bioprocessing - current achievements and future perspectives. *Biotechnol. Appl. Biochem.* **55**, 175-189, (2010).
89 Omasa, T., Onitsuka, M. & Kim, W. D. Cell Engineering and Cultivation of Chinese Hamster Ovary (CHO) Cells. *Curr. Pharm. Biotechnol.* **11**, 233-240, (2010).
90 Josse, L., Smales, C. M. & Tuite, M. F. Transient Expression of Human TorsinA Enhances Secretion of Two Functionally Distinct Proteins in Cultured Chinese Hamster Ovary (CHO) Cells. *Biotechnol. Bioeng.* **105**, 556-566, (2010).
91 Lombrana, M., Moralejo, F. J., Pinto, R. & Martin, J. F. Modulation of *Aspergillus awamori* thaumatin secretion by modification of bipA gene expression. *Appl. Environ. Microbiol.* **70**, 5145-5152, (2004).
92 Moralejo, F. J., Watson, A. J., Jeenes, D. J., Archer, D. B. & Martin, J. F. A defined level of protein disulfide isomerase expression is required for optimal secretion of thaumatin by *Aspergillus awamori*. *Mol. Genet. Genomics* **266**, 246-253, (2001).
93 Messens, J. & Collet, J. F. Pathways of disulfide bond formation in Escherichia coli. *International Journal of Biochemistry & Cell Biology* **38**, 1050-1062, (2006).
94 Fernandez, L. A. & de Lorenzo, V. Formation of disulphide bonds during secretion of proteins through the periplasmic-independent type I pathway. *Mol. Microbiol.* **40**, 332-346, (2001).
95 Lu, Q. *et al.* Optimized procedures for producing biologically active chemokines. *Protein Expression Purif.* **65**, 251-260, (2009).
96 Gasser, B., Maurer, M., Gach, J., Kunert, R. & Mattanovich, D. Engineering of *Pichia pastoris* for improved production of antibody fragments. *Biotechnol. Bioeng.* **94**, 353-361, (2006).
97 van Boxel, G. I. *et al.* Some lessons from the systematic production and structural analysis of soluble [alpha][beta] T-cell receptors. *J. Immunol. Methods* **350**, 14-21, (2009).
98 Nakamura, Y., Gojobori, T. & Ikemura, T. Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Res.* **28**, 292-292, (2000).
99 Vrancken, K. *et al.* pspA overexpression in *Streptomyces lividans* improves both Sec- and Tat-dependent protein secretion. *Appl. Microbiol. Biotechnol.* **73**, 1150-1157, (2007).
100 Shin, H. D. & Chen, R. R. Extracellular Recombinant Protein Production From an *Escherichia coli lpp* Deletion Mutant. *Biotechnol. Bioeng.* **101**, 1288-1296, (2008).
101 Fu, X.-Y. Extracellular accumulation of recombinant protein by *Escherichia coli* in a defined medium. *Appl. Microbiol. Biotechnol.* **88**, 75-86, (2010).
102 Idiris, A. *et al.* Enhanced protein secretion from multiprotease-deficient fission yeast by modification of its vacuolar protein sorting pathway. *Appl. Microbiol. Biotechnol.* **85**, 667-677, (2010).
103 van Wezel, G. P. *et al.* Unlocking *Streptomyces* spp. for use as sustainable industrial production platforms by morphological engineering. *Appl. Environ. Microbiol.* **72**, 5283-5288, (2006).

104 Driouch, H., Roth, A., Dersch, P. & Wittmann, C. Optimized bioprocess for production of fructofuranosidase by recombinant *Aspergillus niger*. *Appl. Microbiol. Biotechnol.* **87**, 2011-2024, (2010).
105 Zhong, J. J. Recent advances in bioreactor engineering. *Korean J. Chem. Eng.* **27**, 1035-1041, (2010).
106 Hewitt, C. J. & Nienow, A. W. in *Adv. Appl. Microbiol.* Vol. 62 105-135 (Elsevier Academic Press Inc, 2007).
107 Shiloach, J. & Fass, R. Growing E-coli to high cell density - A historical perspective on method development. *Biotechnol. Adv.* **23**, 345-357, (2005).
108 Fang, Y., Huang, X. J., Chen, P. C. & Xu, Z. K. Polymer materials for enzyme immobilization and their application in bioreactors. *BMB Rep.* **44**, 87-95, (2011).
109 Abbas, A. K. & Janeway, C. A. Immunology: Improving on Nature in the Twenty-First Century. *Cell* **100**, 129-138, (2000).
110 Zhao, J. R. *et al.* Soluble MHC I and Soluble MIC Molecules: Potential Therapeutic Targets for Cancer. *Int. Rev. Immunol.* **30**, 35-43, (2011).
111 Fernandez, L. A., Sola, I., Enjuanes, L. & de Lorenzo, V. Specific secretion of active single-chain Fv antibodies into the supernatants of *Escherichia coli* cultures by use of the hemolysin system. *Appl. Environ. Microbiol.* **66**, 5024-5029, (2000).
112 Rakestraw, J. A., Sazinsky, S. L., Piatesi, A., Antipov, E. & Wittrup, K. D. Directed evolution of a secretory leader for the improved expression of heterologous proteins and full-length antibodies in *Saccharomyces cerevisiae*. *Biotechnol. Bioeng.* **103**, 1192-1201, (2009).
113 Kjaerulff, S. & Jensen, M. R. Comparison of different signal peptides for secretion of heterologous proteins in fission yeast. *Biochem. Biophys. Res. Commun.* **336**, 974-982, (2005).
114 Proba, K., Worn, A., Honegger, A. & Pluckthun, A. Antibody scFv fragments without disulfide bonds made by molecular evolution. *J. Mol. Biol.* **275**, 245-253, (1998).
115 Desiderio, A. *et al.* A semi-synthetic repertoire of intrinsically stable antibody fragments derived from a single-framework scaffold. *J. Mol. Biol.* **310**, 603-615, (2001).
116 Malik, A., Jenzsch, M., Lubbert, A., Rudolph, R. & Sohling, B. Periplasmic production of native human proinsulin as a fusion to *E. coli* ecotin. *Protein Expression Purif.* **55**, 100-111, (2007).
117 Kjeldsen, T. *et al.* Expression of insulin in yeast: the importance of molecular adaptation for secretion and conversion. *Biotechnol Genet Eng Rev* **18**, 89-121, (2001).
118 Kjeldsen, T. Yeast secretory expression of insulin precursors. *Appl. Microbiol. Biotechnol.* **54**, 277-286, (2000).
119 Cornelis, G. R. The type III secretion injectisome: a complex nanomachine for intracellular 'toxin' delivery. *Biol. Chem.* **391**, 745-751, (2010).
120 Galan, J. E. The *Salmonella* typhimurium type III protein secretion system: an effective antigen delivery platform for cancer therapeutics. *Drugs of the Future* **32**, 985-990, (2007).

121 Garmory, H. S., Brown, K. A. & Titball, R. W. *Salmonella* vaccines for use in humans: present and future perspectives. *FEMS Microbiol. Rev.* **26**, 339-353, (2002).
122 Moreno, M., Kramer, M. G., Yim, L. & Chabalgoity, J. A. *Salmonella* as Live Trojan Horse for Vaccine Development and Cancer Gene Therapy. *Curr. Gene Ther.* **10**, 56-76, (2010).
123 Fronzes, R., Christie, P. J. & Waksman, G. The structural biology of type IV secretion systems. *Nature Reviews Microbiology* **7**, 703-714, (2009).
124 Jong, W. S. P., Sauri, A. & Luirink, J. Extracellular production of recombinant proteins using bacterial autotransporters. *Curr. Opin. Biotechnol.* **21**, 646-652, (2010).
125 Klemm, D., Heublein, B., Fink, H. P. & Bohn, A. Cellulose: Fascinating biopolymer and sustainable raw material. *Angew. Chem.-Int. Edit.* **44**, 3358-3393, (2005).
126 Kaur, J., Chadha, B. S., Kumar, B. A. & Saini, H. S. Purification and characterization of two endoglucanases from *Melanocarpus* sp MTCC 3922. *Bioresour. Technol.* **98**, 74-81, (2007).
127 Kitagawa, T. *et al.* Identification of genes that enhance cellulase protein production in yeast. *J. Biotechnol.* **151**, 194-203, (2011).
128 Handke, P., Lynch, S. A. & Gill, R. T. Application and engineering of fatty acid biosynthesis in *Escherichia coli* for advanced fuels and chemicals. *Metab. Eng.* **13**, 28-37, (2011).
129 Yan, Y. & Liao, J. Engineering metabolic systems for production of advanced fuels. *J. Ind. Microbiol. Biotechnol.* **36**, 471-479, (2009).
130 Lennen, R. M., Braden, D. J., West, R. M., Dumesic, J. A. & Pfleger, B. F. A Process for Microbial Hydrocarbon Synthesis: Overproduction of Fatty Acids in *Escherichia coli* and Catalytic Conversion to Alkanes. *Biotechnol. Bioeng.* **106**, 193-202, (2010).
131 Schirmer, A., Rude, M. A., Li, X., Popova, E. & del Cardayre, S. B. Microbial Biosynthesis of Alkanes. *Science* **329**, 559-562, (2010).
132 Tan, X. M. *et al.* Photosynthesis driven conversion of carbon dioxide to fatty alcohols and hydrocarbons in cyanobacteria. *Metab. Eng.* **13**, 169-176, (2011).
133 Kluge, J. A., Rabotyagova, U., Leisk, G. G. & Kaplan, D. L. Spider silks and their applications. *Trends Biotechnol.* **26**, 244-251, (2008).
134 Widmaier, D. M. *et al.* Engineering the *Salmonella* type III secretion system to export spider silk monomers. *Mol Syst Biol* **5**, (2009).
135 Graf, A., Dragosits, M., Gasser, B. & Mattanovich, D. Yeast systems biotechnology for the production of heterologous proteins. *FEMS Yeast Res.* **9**, 335-348, (2009).
136 Park, J. H., Lee, S. Y., Kim, T. Y. & Kim, H. U. Application of systems biology for bioprocess development. *Trends Biotechnol.* **26**, 404-412, (2008).
137 Hazes, B. & Frost, L. Towards a systems biology approach to study type II/IV secretion systems. *Biochim. Biophys. Acta-Biomembr.* **1778**, 1839-1850, (2008).
138 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29, (2000).

139 Wren, J. D. & Bateman, A. Databases, data tombs and dust in the wind. *Bioinformatics* **24**, 2127-2128, (2008).
140 Gasser, B., Dragosits, M. & Mattanovich, D. Engineering of biotin-prototrophy in Pichia pastoris for robust production processes. *Metab. Eng.* **12**, 573-580, (2010).