Altering the solubility of recombinant proteins through modification of surface features

A thesis submitted to the University of Manchester
for the degree of Doctor of Philosophy
in the Faculty of Life Sciences

2014

MANUEL ALEJANDRO CARBALLO AMADOR

Table of Contents

List of figures

Chapter 1 Figures.

Chapter 2 Figures.

Chapter 3 Figures.

Chapter 4 Figures.

Chapter 5 Figures.

Chapter 6 Figures.

Chapter 9 Figures.

List of Tables

Chapter 1 Tables.

Chapter 2 Tables.

Chapter 3 Tables.

Chapter 4 Tables.

Chapter 5 Tables.

Chapter 6 Tables.

Chapter 6 Tables.

Word count: 48471

Abstract

Institution: The University of Manchester
Name: Manuel Alejandro Carballo Amador
Degree Title: PhD Biotechnology
Thesis Title: Altering the solubility of recombinant proteins through modification of surface features
Date: 2014

Protein solubility plays an important role whether for biophysical and structural studies, or for production and delivery of therapeutic proteins. Poor solubility could lead to protein aggregation, which is an undesired physicochemical mechanism at any stage of recombinant proteins production. To date, more than half of all recombinant therapeutic proteins are produced in mammalian cells, mainly due to the high similarity of the final product to human protein structures. However, poor secretion can occur, due to misfolded proteins or aggregates leading to cellular stress and proteolysis. Another widely-used expression system is *E. coli*, which can offer a cost-efficient alternative. This system has an important limitation, since proteins tends to form insoluble protein aggregates in the cytoplasm upon heterologous overexpression. Several strategies are being implemented to improved soluble expression, ranging from culture conditions to solubility enhancing tags. However, there is no universal approach or technology that solves protein aggregation.

In this thesis two recently published hypotheses from our group have been applied. One stated that soluble expression of proteins was inversely correlated with the size of the largest positively-charged patch on the protein surface. The second hypothesis (of protein solubility), arose from the finding that the relative content of lysine and arginine residues separated *E. coli* proteins by solubility. Both hypotheses arose from a study of an extensive dataset of experimental solubilities determined for cell-free expression of *E. coli* proteins. In combination with other widely used strategies, such as lowering expression temperature and inducer concentration, decreasing non-charged (hydrophobic) patches and addition of helical capping for increasing stability, a rational understanding for directed alteration of solubility in a variety of recombinant proteins has been explored. This includes three protein models to test: (i) recombinant human erythropoietin (rHuEPO) (one of the top selling therapeutics) (ii) recombinant 6-Phosphofructo-2-Kinase/fructose-2,6-bisphosphatase (rPFKFB3) (a product for which over-expression has been sought for characterisation and insight into possible cancer therapy) and (iii) a set of three selected *E. coli* proteins containing high ratios of lysines to arginines: thioredoxin-1 (TRX), cold shock-like protein cspB (cspB), and the histidine-containing phosphocarrier protein (HPr).

It was found that single or multiple point mutations (changing amino acids from positive to negative charge or *vice versa*; or lysines to arginines) verified the predicted effect on rHuEPO, rPFKFB3, TRX, cspB, and HPr solubility (experimentally defined as the distribution between soluble and total fractions) for expression in *E. coli*. In addition, the redesigned set of rHuEPO transiently expressed in HEK 293-EBNA cells, suggesting that positively-charged patch size may also influence protein secretion. Further application of these computational and experimental approaches could provide a valuable tool in the design and engineering of proteins, with enhanced solubility, stability and secretion.

Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright statement

List of abbreviations

aa	Amino acid
Ad5	Human adenovirus type 5
ADP	Adenosine diphosphate
AhpC	Alkyl hydroperoxide peroxidase subunit C
apoE	Apolipoprotein E
ATP	Adenosine triphosphate
AUC	Analytical ultracentrifugation
B. activity	Biological activity
B-factors	Debye–Waller factor or temperature factor
BHK	Baby Hamster Kidney fibroblasts
BiP	Binding immunoglobulin protein or GRP78
BLAST	Basic Local Alignment Search Tool
CD	Circular dichroism
cDNA	Complementary DNA
CDR	Complementarity determining region
CHO	Chinese hamster ovary
CHOP	C/EBP homologous protein or GADD153
CMV	Human cytomegalovirus
cspB	Cold shock-like protein cspB
Ctrl	Control
DB	Disulphide bond
DLS	Dynamic light scattering
DMEM	Dulbecco's Modified Eagle's Medium
DNA	Deoxyribonucleic acid
DSLS	Differential static light scattering
DTT	Dithiothreitol
EBNA	Epstein-Barr virus nuclear antigen
EBV	Epstein-Barr Virus
EC	Enzyme Commission number
EDTA	Ethylenediaminetetraacetic acid
EGFR	Epidermal growth factor receptor
EmGFP	Emerald Green Fluorescent Protein
EOR	ER overload response
EPO	Erythropoietin
ER	Endoplasmic reticulum
ERAD	ER associated degradation
ERK	Extracellular signal regulated kinase
eSOL	The solubility database of all *E. coli* proteins
F6P	Fructose-6-phosphate
For	Forward primer
Fru-2,6-P_2	Fructose-2,-bisphosphate
G-CSF	Granulocyte colony-stimulating factor
GFP	Green fluorescent protein
gor	Gene encoding glutathione reductase
h	Hour
HEK	Human embryonic kidney cells
hGH	Human growth hormone

HIV-1	Human immunodeficiency virus type 1
HPCE	High performance capillary electrophoresis
HPr	Histidine-containing phosphocarrier protein
HSP	Heat shock protein
HuEPO	Human EPO
Hyd	Hydrophobic
IBs	Inclusion bodies
Ig	Immunoglobulin
IPTG	Isopropyl β-D-1-thiogalactopyranoside
IR	Infrared
kDa	Kilodalton
LB	Luria-Bertani broth
MALDI-TOF	Matrix-assisted laser desorption ionization- time-of-flight mass spectrometer
MBP	Maltose-binding protein
mRNA	Messenger RNA
MS	Mass spectrometry
n/a	Not applicable
NADH	Nicotinamide adenine dinucleotide
Neg	Negative
NMR	Nuclear magnetic resonance
nonQ	Non-charge patch
nonQmax	The maximal size of a non-charged patch
NusA	N-utilisation substance A
OD	Optical density
ORF	Open reading frame
oriP	Replication origin
P. pastoris	*Pichia pastoris*
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
PDB	Protein Data Bank
PDI	Protein disulphide isomerase
PEP	Phosphoenolpyruvate
PFKFB3	6-Phosphofructo-2-Kinase/fructose-2,6-bisphosphatase or iPFK2
pI	Isoelectric point
pLysS	T7 lysozyme coding sequence
Pos	Positive
posQ	Positive patches
posQmax	The maximal positively charged patch
PSI-BLAST	Position-Specific Iterated-BLAST
PTMs	Post-translational modifications
PTS	Carbohydrate phosphotransferase system
QC	Quality control
Rev	Reverse primer
RevW-H	Revised Wilkinson-Harrison predictor
rHuEPO	Recombinant HuEPO
RNA	Ribonucleic acid
RNP-1/2	RNA recognition motif
rPFKFB3	Recombinant PFKFB3
s	Seconds

SAP	Spatial aggregation propensity
SDM	Site-directed mutagenesis
SDS-PAGE	Sodium dodecyl sulphate-polyacrylamide gel electrophoresis
SEC	Size exclusion chromatography
SEM	Standard error of the mean
SI DHFR	Selectivity index dihydrofolate reductase
Sol	Soluble
Sup	Supernatant
Tagg	Aggregation onset temperature
TBS	Tris-buffered saline
TF-1	Human erythroleukaemia cells
TGN	*trans*-Golgi network
TM	Transmembrane domain
Tm	Unfolding transition temperature
tPA	Tissue plasminogen activator
TRiC	TCP-1 Ring Complex
tRNA	Transfer RNA
TRX	Thioredoxin-1
trxB	Gene encoding thioredoxin reductase
U	Units
UPR	Unfolded protein response
v/v	Volume/volume
w/v	Weight/volume
WT	Wild-type
XBP1	X-box binding protein 1
Δ	Deletion/replacement

Acknowledgements

Firstly, I would like to express my deepest gratitude to my supervisors Prof. Alan Dickson and Dr. Jim Warwicker for all their invaluable support and guidance throughout my PhD. Both opened up their time and space to welcome me from another part of the world, it is a true privilege to be part of their groups. I will always be grateful for the opportunity to work and learn from their great values and the friendly environment. Additionally, I would like to extend my gratitude to my advisor Prof. Stephen High for the very useful discussions throughout my project. Also, I would like to thank Dr. Edward McKenzie for sharing his help, materials, expertise and enjoyable coffee moments.

Secondly, I would like to acknowledge my colleagues and friends in the Dickson Lab and Warwicker group. Thanks for providing a friendly environment that it made easier to work during sunny and rain days alike.

I would also like to thank all the wonderful people that I have met during my PhD, especially to my roommates for being like a family to me in this part of the world. Thanks to "Supersonicos" for all those amazing years of fun and friendship. Thanks to "The overground old" for the music and fun. Thanks to Dr. Stephano for all your unconditional support and guidance in my scientific career. Also, thanks to all my dear friends back home who supported me and believed in me, especially to Luis, Robert and Liz.

I would like to gratefully acknowledge CONACyT for funding.

I would like to dedicate this thesis to my family for their never-ending love and support, especially to my Mum, Grandma, sister and uncle. It is also dedicated to Liss, who encouraged me to start this journey, thanks for all your unconditional support and love for all these beautiful years.

"As life's language is comprised of the combination of four letters in carbon ink, synthetic or artificial intelligence is based on two numbers written in silicon dreams"

Jando

Chapter 1

Introduction

1.1 Overview

Life forms are encrypted with their own genetic code, which carry the necessary information to sustain life. The translation of this information results in the synthesis of a wide range of polypeptides, which encompass an extensive variety of properties as a result of millions of years of evolutionary pressure. A single cell has a complex environment comprising abundant mechanisms to maintain fitness. These mechanisms can be unbalanced with the imposition of recombinant technologies. The insertion of foreign genes, whether for therapeutics or research purposes, gives new challenges for the cell. Overexpression of a target protein can overload cellular machinery, often leading to protein aggregation. This has been a central topic for decades in biotechnology and industry (Ventura, 2005, Chan et al., 2013). Several attempts to understand the molecular events that lead to aggregation upon heterologous expression have been made using approaches that utilise genetics and cellular and/or protein engineering. Over the past decades with the growth in biophysical understanding of proteins and enhanced computational technology, several computational approaches have been developed to understand and offer prediction of molecular features and environmental conditions that influence protein aggregation. The theme of this thesis is to enhance understanding of predictive approaches that use the biophysical characterisation of organised proteins structures to alter the solubility of recombinant proteins, with particular focus on the application of computational algorithms developed recently in our group.

1.2 Recombinant protein expression

The universality of the genetic code and progress in biotechnology have enabled the development of technologies for the production of foreign proteins in heterologous expression systems. Since the launch of first successful recombinant proteins used in clinical practice in the seventies (somatostatin followed by insulin), protein production has been accompanied by experimental challenges and has generated great economic impact (Liras, 2008, Walsh, 2014). Production of highly concentrated and soluble proteins is a necessary requirement for use of proteins for biophysical and structural studies or for therapeutic (biopharmaceutical) purposes (Esposito and Chatterjee, 2006). In order to achieve these requirements, prokaryotic and eukaryotic expression systems are widely used in the production of recombinant proteins (Palomares et al., 2004, Bhopale and Nanda, 2005, Aricescu et al., 2006). An extensive range of host organisms are available as platforms for production/manufacture of recombinant proteins, including bacteria, yeast, fungi, plant cells, microalgae, insect cells and mammalian cell systems (Mahmoud, 2007, Specht et al., 2010). Although there has been great progress in the effectiveness of these expression systems, each has advantages and disadvantages, from low cost with high yields to limitations in yield scale of expression or the profile of post-translational processing that may be obtained (Palomares et al., 2004). The appropriate decision in choice of a specific expression system depends on the nature of the protein of interest, i.e. required scale of production, required post-translational modifications (PTMs), protein mass, number of disulphide bonds (DB), secretion and purification mechanisms (Brondyk, 2009).

1.3 Protein expression systems

Proteins are one of the most profitable products among the current therapeutics for commercial purposes (Pavlou and Reichert, 2004, Walsh, 2014). In order to fulfil the demand, important considerations need to be made in the choice of expression system, in order to select the most cost-efficient production (Table 1.1). The bulk of the drug market is destined for human consumption. Hence, mammalian cells are ideal due to their capabilities to direct protein folding, secretion and complex PTMs (Wurm, 2004). An efficient procedure along the secretory pathway results in an ideal final product, since most of the commercial therapeutics are secreted proteins (e.g. hormones, interferon, monoclonal antibodies) (Peng and Fussenegger, 2009).

Escherichia coli may offer several benefits for specific protein products, which range from rapid biomass accumulation and inexpensive manipulation to a straightforward scale-up (Baneyx and Mujacic, 2004). Despite these benefits, *E. coli* has an important drawback, it is unable to carry out PTMs such as occurs in eukaryotic cells. However, there are several examples of therapeutics that do not require complex PTMs to be effective medicines, such as insulin, human growth hormone, interferon -α, -β and -γ, interleukin-2, and tumour necrosis factor -α (Walsh and Jefferis, 2006, Arya et al., 2008, Dingermann, 2008). In addition, Jeong et al. have recently shown that recombinant human erythropoietin (rHuEPO) expressed in *E. coli* does not require glycosylation in order to be active in an *in vitro* proliferation assay (Jeong et al., 2014).

Table 1.1. Features of expression system for recombinant therapeutic proteins production.

Expression system	PTMs			Production		
	Disulphide bonds	**Glycosylation**	**Secretion**	**Costs of fermentation**	**Safety costs**	**Processes developed**
Bacterial (*E. coli*)	✔ (In the periplasm)	✖	Periplasmic secretion	Promoter-dependent low to moderate	Low	Industrial scale
Yeast (*P. pastoris*)	✔	✔ No terminal α1,3 mannose	Possible	Low	Low	Industrial scale
Plant cells	✔	✔ Terminal fucose	Possible; size-restriction	Moderate	Low	Pilot scale; production scale
Mammalian cells (e.g. CHO)	✔	✔ (typically human-like)	Usually	High	High	Industrial scale
Animals (Mammals)	✔	✔ (typically human-like)	Usually	Moderate (Farming)	High	Industrial scale

Adapted from (Dingermann, 2008).

1.3.1 *Escherichia coli* as a protein expression system

One of the main limitations in this expression platform, whether for biophysical studies or industrial-scale is the formation of insoluble protein aggregates (inclusion bodies, IBs) in the cytoplasm or periplasm (Baneyx and Mujacic, 2004). IB formation is associated with incomplete or incorrect folded proteins, usually upon heterologous overexpression (e.g. by use of strong promoters and high inducer concentration), linked to production rate surpassing the capacity of folding modulators to handle the protein (Baneyx and Mujacic, 2004).

Consequently, for harvesting such proteins for subsequent use, extra processing is required to achieve protein solubilisation by recovering, refolding and purification. For industrial purposes, this presents risks since each protein (with specific biophysical characteristics) may demand development of specific methods for solubilisation. Costs and timelines for solubilisation are not always suitable in industrial-scale processes (Ventura and Villaverde, 2006). Also, solubilisation may diminish not only the amount of the target protein but also the bioactivity. For example, human growth hormone (hGH) tends to form IBs when expressed in the cytoplasm, whereas when targeted to the periplasm it has been recovered as soluble and bioactive protein (Sockolosky and Szoka, 2013). *E. coli* offers different expression compartments such as cytoplasm, periplasm, cell surface, and secretion to the medium (Cornelis, 2000). Nevertheless, cytoplasmic and periplasmic expression are widely used for production of soluble proteins. Expression in the cytoplasm is generally chosen due to the high yield production (Sørensen and Mortensen, 2005a). However, the oxidative environment in the periplasm allows disulphide bond formation (Depuydt et al., 2009), whereas the reducing environment of the cytoplasm does not (Lobstein et al., 2012). To solve this drawback, a commercial available strain has been engineered, SHuffle. This strain, among the wide *E. coli* engineered repertoire (Table 1.2), is characterised by the absence of two reductases (thioredoxin reductase [*trxB*] and glutathione reductase [*gor*]) but the presence of a copy of the periplasmic DsbC disulphide bond isomerase in the cytoplasm (Lobstein et al., 2012). This environment favours the folding and disulphide bond formation of the overexpressed recombinant proteins (Fig. 1.1). In addition to these benefits, translocation of recombinant proteins into the periplasm also allow less protein degradation, it contains 4% of the total cell protein (~100 different proteins) and easy purification by osmotic shock (Jonasson et al., 2002, Schumann and Ferreira, 2004). Despite these advantages, some proteins when directed to the periplasm resulted in incomplete translocation through the membrane or IB formation when

overloading the machinery capacity leading to degradation in the cytoplasm (Baneyx, 1999), and protein leakage to the extracellular cultivation medium (Sørensen and Mortensen, 2005a).

Table 1.2. Features of engineered *E. coli* strains common used for heterologous protein expression.

***E. coli* strain**	**Aim**	**Features**
BL21	Less protease degradation of recombinant protein	Deficient in proteases *Lon* and *OmpT*
BL21-CodonPlus(DE3)	Overcome the effect of codon biasness	tRNA genes: *argU*, *ileY* , *proL* and *leuW*
BL21 (DE3)	T7 Expression Strain	T7 polymerase encoded
BL21 (DE3) pLysS	Controlled expression	Lysozyme encoded plasmid
Origami	Enhance disulphide bond formation in the cytoplasm	*gor* and *trxB* reductases genes mutated
Rosetta	Both the AT and GC rich gene	All the rare tRNA coding gene
SHuffle	Proper disulphide bond formation	*gor* and *trxB* reductases genes mutated and DsbC gene encoded

Adapted from (Gopal and Kumar, 2013)

Targeting of recombinant proteins to the periplasmic compartment could involve two major well-studied systems (Fig. 1.1), the general secretion pathway (Sec pathway) or the twin-arginine translocation (Tat pathway) (Thomas et al., 2001). In both systems, proteins need to encompass a signal-sequence at their N-terminal (15-30 amino acids long) in order to be translocated. Many heterologous proteins with a Sec-dependent signal peptide fused have failed due to folding considerations (Lee et al., 2006). These are proteins with prolonged hydrophobic areas (captured within the membrane) and those which fold rapidly within the cytoplasm (Schumann and Ferreira, 2004). These drawbacks could be overcome using the Tat

pathway, since one of the advantages is the ability to export folded proteins (Robinson and Bolhuis, 2001, DeLisa et al., 2003). Despite the widely-studied Sec pathway, several recombinant proteins are not compatible with the mechanism involved, therefore the Tat pathway seems promising for biotechnological purposes (Lee et al., 2006). However, many eukaryotic proteins of commercial or therapeutic importance encompass complex structures and frequently involve disulphide bonds and PTMs (Baneyx and Mujacic, 2004). Based on these challenges in *E. coli* expression, mammalian cells are at present the widely used system for production of recombinant proteins in the pharmaceutical industry.

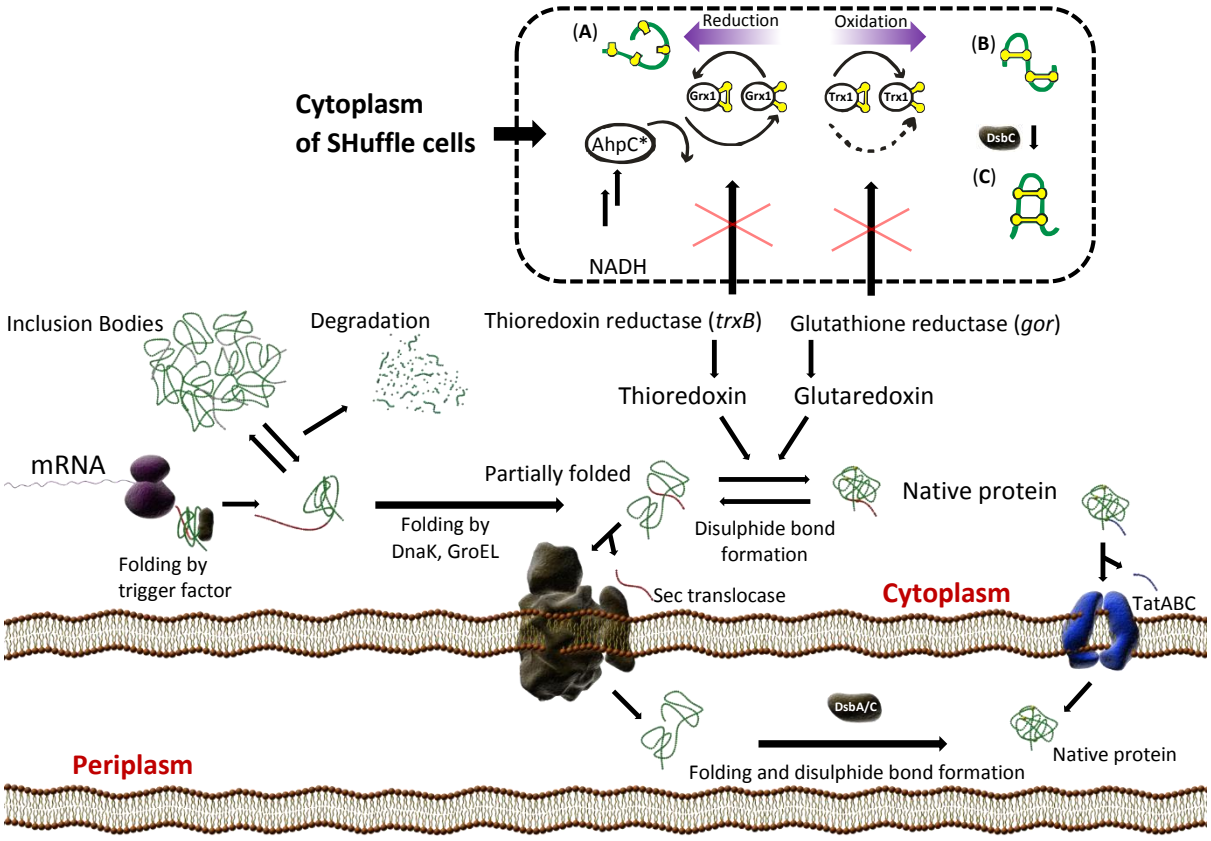


Fig. 1.1. Protein translation, folding and secretion in *E. coli*. As soon as the nascent polypeptides leave the exit tunnel of the *E. coli* ribosome they associate with the trigger factor chaperone and the folding mechanisms are initiated. After release from trigger factor, proteins can continue to fold into their native state. Those that are trapped in non-native (or partially

folded) states, or in a conformation prone to aggregation, are targets for DnaK and GroEL chaperones. Chaperones may prevent IB formation by diminishing aggregation and promoting degradation of misfolded proteins, or by mediating the solubilisation or disaggregation of proteins. Targeting proteins to the periplasmic space, whether by Sec or Tat pathway, can be beneficial for disulphide bond formation. This is also possible in the engineered cytoplasm of SHuffle strain (*Δgor ΔtrxB*). (A) Reduced protein by Grx1 or oxidised by Trx1. (B) Mis-oxidised protein is isomerised to its native correctly folded state (C) by DsbC. Redox states of cysteines are indicated as yellow circles (oxidised = circle + stick; reduced = circle). Adapted from Baneyx and Mujacic, 2004, Sørensen and Mortensen, 2005, and Lobstein et al., 2012.

1.3.2 Mammalian cells as a protein expression system

To date, mammalians cells are generally used in the recombinant biopharmaceutical industry, since they can undergo the different PTMs that are often encompassed for protein activity and stability (Le Fourn et al., 2014). Mammalian cell expression system can be developed for transient (not integrated into the genome) or stable expression (gene integrated in the genome) depending upon the purpose or the features of the heterologous protein. Several studies to improve recombinant expression have been developed over the past decades, which focused on translational or secretory capacity of mammalian host cells (Barnes and Dickson, 2006). The secretion capacity of a host cell is considered the major bottleneck after reached an apparent limit in transcription and translational engineering strategies (Peng and Fussenegger, 2009). Since most of the therapeutic proteins are secreted (e.g. hormones or monoclonal antibodies) understanding protein features along the secretory pathway leads to an approachable area to improve recombinant protein production.

Protein secretion takes effect through a series of subcellular compartments. Newly synthesized polypeptides possessing an N-terminal hydrophobic signal peptide are translocated *via* different pathways (e.g. the universally conserved co-translational pathway) into the lumen of endoplasmic reticulum (ER) (Nyathi et al., 2013). In the ER proteins undergo folding and then continue to the Golgi apparatus where they are processed and packaged to be released to the extracellular space (Palade, 1975). This is only a simplified representation. In theory there occurs two types of secretion in cells: constitutive and regulated (Fig. 1.2). The former one consists in secreting proteins as soon as these are synthesized. Regulated secretion first stores nascent proteins in transport vesicles before being secreted (Peng and Fussenegger, 2009). Different compartments along the secretory pathway involve different environments, such as pH, ranging from 7.2 in the ER lumen to 5.2 in the secretory granules (Paroutis et al., 2004).

After translation, the polypeptide chain is processed through different PTMs in separate subcellular compartments, such as cytosol, ER and Golgi apparatus (Blom et al., 2004). Post-translational modifications occur naturally during the protein biosynthesis and may take place throughout or after translation (Chung et al., 2011). Some of these modifications are associated with therapeutic proteins, which mainly include glycosylation, disulphide bond formation, asparagine deamidation, methionine oxidation and proteolysis (Walsh and Jefferis, 2006). In the therapeutic context, glycoproteins may present N-linked glycosylation and/or O-linked glycosylation (Solá and Griebenow, 2010). The activity of some recombinant drugs are determined by their structure, for some others glycosylation shows an essential role in the function (e.g. protection from proteolysis, solubility or receptor binding) (Werner et al., 2007). Adding extra glycosylation sites can impart several benefits, which include protection from unfolding and denaturation (by pH, heat, chemical and freezing), protein solubility, protein targeting/trafficking, ligand recognition/binding, biological activity, half-life, cross-linking

and immunogenicity (Kobata, 1992, Willey, 1999, Elliott et al., 2003, Solá and Griebenow, 2009).

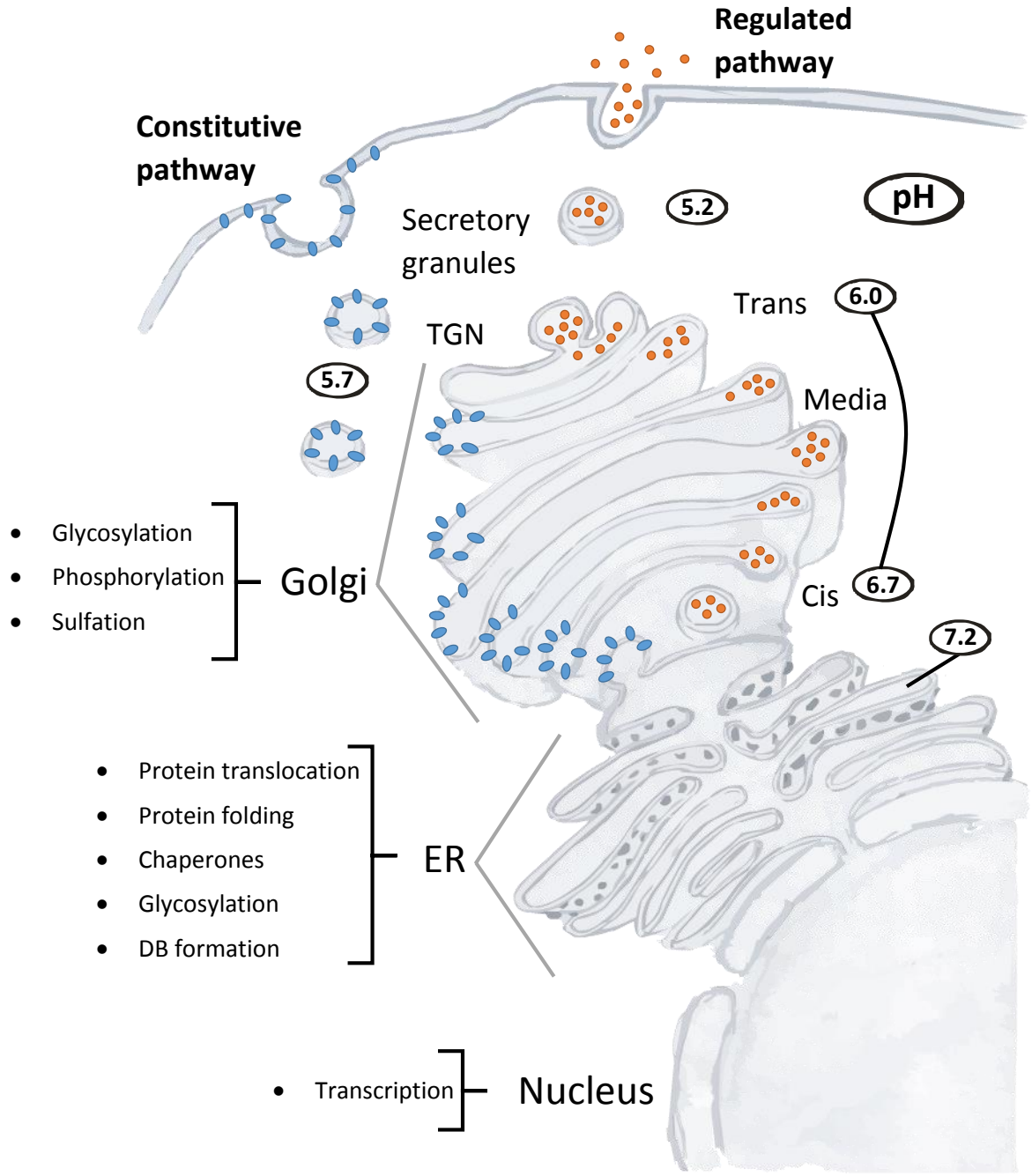


Fig. 1.2. Comparison of the steady-state pH of the compartments of secretory pathways. Overview of the recombinant protein expression along the secretory pathway in mammalian cells. TGN, *trans*-Golgi network. Adapted from Paroutis et al., 2004. See text for details and references.

Another important PTM is disulphide bond formation, which arises from the covalent bond formed by the oxidation of sulfhydryl groups between cysteine side chains (Lobstein et al., 2012). This represents a crucial component for stability in some multi-chain proteins, such as immunoglobulins (Jenkins et al., 2008). Correct disulphide bond formation is vital for the stability and folding of many secreted proteins (Frand et al., 2000). In eukaryotic cells, DB formation occurs in the oxidative environment of ER lumen that has abundant calcium, favouring a proper protein folding (Costa et al., 2011). This mechanism is carried out by an enzymatic reaction between the protein disulphide isomerase (PDI) with the nascent peptide chain (Freedman, 1989). Mohan et al., showed by transfecting more copies of the PDI gene into Chinese hamster ovary (CHO) cells, an increase from 15 to 27% in antibody productivity was achieved (Mohan et al., 2007). In contrast, poor secretion may be due to misfolded proteins or aggregates leading to triggering of the ER stress and the unfolded protein response (UPR) (Le Fourn et al., 2014).

1.3.2.1 Quality Control in the secretory pathway: Protein degradation

Quality control (QC) comprises mechanisms that identify non-functional or altered intracellular elements and eliminate them inside the cell (Ellgaard and Helenius, 2003). These mechanisms involve chaperones that recognise and target misfolded proteins to degradation. Cells depend on the proper function of QC machinery to carry out their vital functions (Cuervo et al., 2010). In this context, misfolding proteins undergo proteolysis in the UPR or by the ER associated protein degradation (ERAD) pathway (Fig. 1. 3). Unfolded or aggregated proteins saturate the folding and/or processing capacity of the ER, which leads to ER stress. This situation triggers the UPR in order to cope with the accumulation of unfolded or misfolded

proteins (Fig. 1.3) (Schröder and Kaufman, 2005). In addition, a second ER response has been described to deal with the excess of proteins: ER overload response (EOR). EOR acts on excess accumulation of proteins in the ER membrane and UPR on overload of unfolded proteins in the ER lumen (Cudna and Dickson, 2003). Understanding and solving these responses along the secretory pathway may lead to a better yield production in commercial based host cells (e.g. CHO, NS0, BHK, HEK-293).

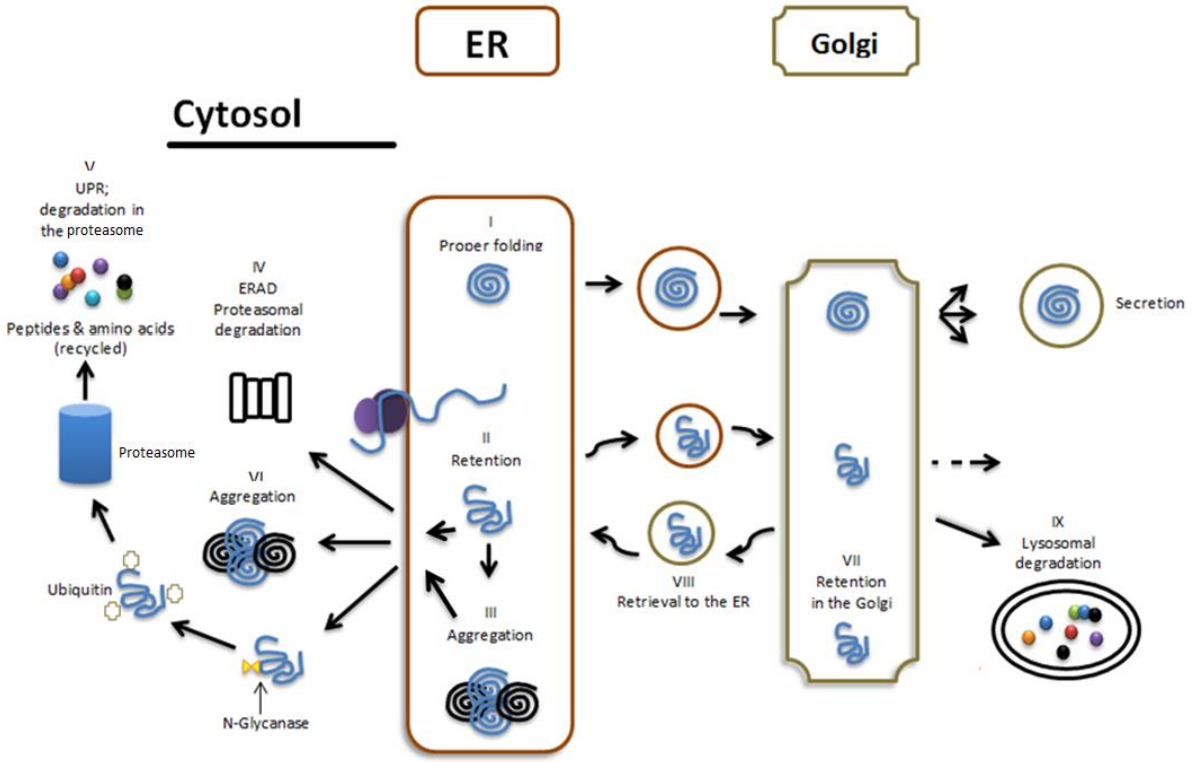


Fig. 1. 3. Summary of the quality control stages along the secretory pathway. Nascent polypeptides are translocated into the ER to undergo structural changes. (I) Properly folded proteins are transported out of the ER. (II) Unfolded or misfolded proteins are generally retained in the ER and exposed to ER resident chaperones to attempt folding. (III) Misfolded proteins may aggregate in the ER (transiently or permanently), or they may be expelled to the cytosol to be processed through the ERAD pathway (IV) or to the UPR to be degraded in the proteasome (V); or aggregation in the cytosol (VI). Certain partly folded proteins are

transported to the Golgi apparatus, where they may be retained (VII), or returned to the ER (VIII), or targeted to lysosomes for degradation (IX). Adapted from (Trombetta and Parodi, 2003, Jenkins, 2007).

1.3.2.2 Human embryonic kidney 293 EBNA

Human embryonic kidney (HEK) cells transformed with sheared fragments of human adenovirus type 5 (Ad5) DNA developed from the well-studied host cell HEK 293 (Thomas and Smart, 2005). HEK 293 cells are widely used as a transient expression system due to high transfection efficiency and their capacity to support protein production. In addition, this cell line offers most of the post-translational folding and modifications related to both mammalian or non-mammalian sources (Thomas and Smart, 2005). The introduction of vectors under the control of the human cytomegalovirus (CMV) promoter, makes efficient utilisation of the host cell's protein translational machinery. This cell line is often used in preliminary screening of potential proteins for stable or large-scale production (Meissner et al., 2001).

Recombinant HEK293 cells, constitutively expressing the Epstein-Barr virus nuclear antigen 1 (EBNA-1), allow episomal replication of vectors encoding an Epstein-Barr Virus (EBV) replication origin (oriP), which permits high replication (Yates et al., 1985). The interaction between EBNA-1 and oriP is essential for DNA replication and distribution of the vector in daughter cells ensuring a high protein expression (Aiyar et al., 1998, Hung et al., 2001). In addition, stable expression in this cell line is possible under long-term antibiotic maintenance (Parham et al., 2001). Whether for transient or stable expression of biopharmaceuticals, a protein must be stable and soluble long-term for release to the market (Caravella and Lugovskoy, 2010).

1.4 Protein solubility

Different extrinsic and intrinsic factors can affect protein solubility, all of which should be considered in protein engineering. Extrinsic factors include pH, temperature, ionic strength and solution composition (Trevino et al., 2008). On the other hand, intrinsic factors are determined primarily by the amino acids on the protein surface (Kramer et al., 2012). Also, these factors are concerns at any stage of small or large-scale protein production, whether for pharmaceutical industry (Baneyx and Mujacic, 2004) or structural studies (Bagby et al., 2001). Two types of poor solubility can be described: *in vitro* or *in vivo*. Low *in vitro* solubility comprises proteins that are expressed, purified and folded that cannot be concentrated at room temperature (Trevino et al., 2008). Low *in vivo* solubility is related to recombinant proteins upon overexpression in cellular systems.

In *Escherichia coli*, several strategies are being implemented to improved soluble expression, which include protein expression at low temperatures and inducer concentration, different cultivation strategies, co-expression of molecular chaperones, fusion of desired proteins with solubility enhancing tags and "rational" site-directed mutagenesis (Sørensen and Mortensen, 2005b). The best strategy available to date has been fusion tags. The complete mechanism of action is unknown. Nevertheless, it has been observed that fusion tags can help to improve protein folding and result in less protein aggregation in the cytoplasm (Esposito and Chatterjee, 2006).

1.4.1 Protein folding and aggregation

Protein folding is a critical process that transforms a polypeptide from its linear chain configuration to a three dimensional structure (Anfinsen, 1973). The 3D conformation is generally thermodynamically stable, and reaches a native-functional shape (Markossian and Kurganov, 2004). This modification takes place in the ER and cytosol, the former one being more complex since it is where the proteins are being modified (e.g. with glycosylation and disulphide bond formation) (Schröder and Kaufman, 2005). The lumen of the ER has an oxidative environment compared to the cytosol, where some enzymes known as "chaperones" assist folding and secretion of proteins (Jenkins et al., 2009). Proteins can adopt diverse structural conformations, determined by the interactions of their amino acids (Dobson et al., 1998). An early step in protein folding is to hide hydrophobic (non-polar or non-charged) regions within the core of the molecule, breaking contact with the surrounding water. Another folding factor is hiding electrostatic interactions, for example hydrogen or disulphide bonds, in the hydrophobic interior (Stevens and Argon, 1999). These structural arrangements help to protect proteins surfaces against aggregation. Protein aggregation is the oligomerisation and/or polymerisation of unfolded peptides and proteins (Tan et al., 2011).

Protein aggregation involves reversible and irreversible reactions (Andya et al., 2003), such as physical aggregation (non-covalent interactions between hydrophobic surfaces) without structure changes, and/or chemical aggregation (oxidation or covalent bonds, e.g. DB formation/exchange). Although these reaction are determined by amino acid sequences, there are also common external factors that impact protein aggregation, which include temperature, pH and protein concentration (Wang, 2005). Protein aggregation is a non-desired physicochemical mechanism for biopharmaceutical companies at any scale (Weiss et al.,

2009). Therefore, investigate the mechanisms involved in protein folding might provide a better understanding in order to develop strategies to solve protein aggregation.

1.4.2 Contribution of molecular chaperones to protein folding

Molecular chaperones have been described as any protein that interacts in the folding or assembly of a different protein without being part of its native structure (Hartl, 1996). The main role of many molecular chaperones is to assist in protein folding and unfolding processes in order to regulate the proteome homeostasis in the cells (Saibil, 2013). During protein folding or when misfolding take place, partial exposure of non-polar (hydrophobic) regions could lead to aggregation. Molecular chaperones are associated to these non-polar sections to suppress aggregation (Walter and Buchner, 2002). The low specificity of chaperones in these interactions confers the ability to assist in appropriate folding of a large variety of polypeptides that differs in sequence and conformation (Walter and Buchner, 2002). Inherent with the mechanism of action of chaperones is the involvement of cycles of ATP binding and hydrolysis to act on non-native protein structures, which facilitates their folding or unfolding (Mayer, 2010). In addition to activity upon completed protein conformations, chaperones also associate with nascent subunits during their assembly process (Saibil, 2013).

The functionality of molecular chaperones in assisting protein folding and proteostasis has a significant impact in the structural evolution of proteins (Kim et al., 2013). Chaperones are classified into different groups according to their sequence homology and were initially named on the basis to their molecular weight (Table 1.3). The major classes of molecular chaperones are included in several families of stress (e.g. heat shock or oxidative stress) proteins, such as the heat-shock proteins (Hsps) Hsp40s, Hsp60s, Hsp70s, HSP90s, Hsp100s and the small Hsps (Fink, 1998, Kim et al., 2013).

Table 1.3. Components and function of the main chaperone systems in bacteria and eukaryotic cells

Chaperones	Functions
Chaperonin system	
HSP60 (also known as CPN60; eukaryotic), GroEL (*E. coli*), TRiC (also known as CCT; eukaryotic), thermosome (archaea) prefoldin (also known as the Gim complex, GimC; archaea and eukaryotic) small Hsps (also known as holding chaperones; prokaryotic and eukaryotic)	Assist protein folding and prevent aggregation
HSP70 system	
DnaK (*E. coli*), Ssa, Ssb (*S. cerevisiae*), BiP (also known as GRP78; ER)	Protein folding and unfolding, disaggregation, stabilisation of extended chains, translocation across organelle membranes, regulation of the heat-shock response, targeting peptides for degradation
HSP90 system	
HptG (*E. coli*), GRP94 (also known as HSP90B1; ER)	Protein binding, stabilisation and maturation of steroid receptors and protein kinases, targeting to proteases, regulation of substrate selection and fate, myosin assembly
HSP100	
ClpA, ClpB, ClpX, HslU (bacteria; mitochondria and chloroplasts), p97, RPT1–RPT6 (eukaryotic)	Protein unfolding, proteolysis, thermotolerance, resolubilisation of aggregates, protein remodelling

BiP, binding immunoglobulin protein; CCT, chaperonin-containing TCP-1 Ring Complex (TRiC); ER, endoplasmic reticulum; GRP78, 78 kDa glucose-regulated protein; HSP, heat shock protein. Adapted from (Hartl, 1996, Saibil, 2013).

In prokaryotes, trigger factor and DnaK associate with newly synthesised polypeptides (Hesterkamp et al., 1996, Teter et al., 1999), whereas GroEL acts together with its GroES co-chaperone post-translationally (Horwich et al., 1993, Deuerling et al., 2003). Eukaryotic cells lack a trigger factor homolog (Frydman, 2001) but its role may be taken on by some cytosolic chaperones that have been shown to interact with ribosome-bound nascent peptides. The chaperone interactions link to action of the nascent chain–associated complex (NAC) (Wang et al., 1995), Hsp/Hsc70 system (Beckmann et al., 1990, Frydman et al., 1994), Hsp90 (Uma et al., 1997), TRiC/CCT (Frydman et al., 1994), and GimC/prefoldin complex (Hansen et al., 1999). These cytosolic complexes/systems enhance protein folding/refolding in the cytosolic environment. In the ER, BiP (hsp70 family) is one of the most abundant chaperones (Zhang

and Kaufman, 2006). It has been ascribed functions such as enhancement of protein folding and translocation (Ailor and Betenbaugh, 1998, Schröder, 2008). PDI (55 kDa) is a second important ER chaperone which operates to impede aggregation and misfolding in proteins (Wilkinson and Gilbert, 2004). The thiol-disulphide oxidoreductase ERp57 is its closest known homologue (Frickel et al., 2004). Several studies have revealed ERp57 to be part of the calreticulin and calnexin chaperone system, promoting the oxidative folding and quality control of newly synthesized glycoproteins in the ER (Ellgaard and Frickel, 2003).

1.5 Profiling protein aggregation

Aggregation prediction and profiling of proteins has been a constant challenge over the past decades (Hamrang et al., 2013). Several computational approaches have been developed to understand and help solve aggregation issues (Hwang and Park, 2008), along with biophysical and biochemical technologies. Some of these analytical technologies have been used routinely as instrumentation for profiling aggregates, including dynamic light scattering (DLS), size exclusion chromatography (SEC), circular dichroism (CD), analytical ultracentrifugation (AUC), and differential scanning fluorimetry. Others have been emerging for this application, such as nuclear magnetic resonance (NMR), mass spectrometry (MS), differential static light scattering (DSLS), and high performance capillary electrophoresis (HPCE) (Hamrang et al., 2013). These techniques offer a useful screening with their own advantages and disadvantages, ranging from small sample requirements and high sensitive to semi-quantitative and limited particle size range, respectively. For example, SEC is a popular profiling method used for detection and quantification of protein aggregates (Mahler et al., 2009). It is highly sensitive, small sample volume is required and is reproducible, however, it is limited to certain particle

sizes and requires sample dilution and filtration (Carpenter et al., 2010, Zölls et al., 2012). The choice of computational approaches and/or biophysical analytical methods depends upon the target protein features.

1.6 Experimental approaches to increase solubility

Several experimental attempts to reduce hydrophobic regions on protein surfaces have been made. Some of these studies include substitutions such as L435K on Moloney murine leukemia virus reverse transcriptase (Das and Georgiadis, 2001), F185K HIV-1 integrase (Jenkins et al., 1995), a five mutations version of Human apoE including V287E (Fan et al., 2004), F185H of SIV integrase (Li et al., 1999), and the C-terminal tetrapeptide FKPY replacement for a single glycine at position 324 of the HhaI methyltransferase (Daujotytė et al., 2003).

A particular approach which involves changing amino acids to negatively charged residues has been emerging to increase solubility. Trevino et al. proved that aspartic acid, glutamic acid, and serine are the most favourable amino acids on the solvent-exposed T76 of ribonuclease Sa, especially at high net charge (Trevino et al., 2007). In addition, more experiments had been made which include the substitutions N48E and N130D of type SI DHFR (Dale et al., 1994), W100E of the obese protein leptin (Zhang et al., 1997) and Human apoE V287E (Fan et al., 2004). Additionally, Gallagher and co-workers demonstrated that mutations D19A or D19N in the class IV adenylyl cyclase from *Yersinia pestis* result in a less soluble protein. In this example, mutants were not able to crystallize (Gallagher et al., 2009). Moreover, multiple structure-based mutations on a single protein are being favoured to increase solubility (Mosavi and Peng, 2003, Slovic et al., 2004, Fowler et al., 2005, Roosild and Choe, 2005).

Some of these approaches are directed to obtain protein crystallization (Das and Georgiadis, 2001), long-term storage and to conserve or increase activity of proteins (Fowler et al., 2005).

Despite the advances in technology and knowledge, there is no universal method to overcome aggregation problems. Combining computational, biophysical and experimental approaches, may help to diminish this problem in order to produce more effective proteins in a favourable period frame.

1.7 Hypothesis and aims

Protein aggregation is related to protein sequence, structure and stability. The work in this thesis focusses mostly, but not entirely, on engineering the native structural properties of a protein, rather than alteration of stability. In particular, modifying amino acids on protein surfaces to diminish positively-charged patches could be associated with improved protein solubility in over-expression, according to a recently published hypothesis (Chan et al 2013). From this basis the theme of this thesis is derived, a theme devoted to alteration of the solubility of recombinant proteins, which include three protein models to test: (i) human erythropoietin (HuEPO) (one of the top selling therapeutics) (ii) 6-Phosphofructo-2-Kinase/fructose-2,6-bisphosphatase (PFKFB3) (a product for which over-expression has been sought for characterisation and insight into possible cancer therapy) and (iii) a set of three selected *E. coli* proteins containing high ratios of lysines to arginines (thioredoxin, cold shock-like protein cspB, and the histidine-containing phosphocarrier protein). These proteins are selected as models for redesign and exploration of two computational approaches that were recently developed in our group. Protein solubility analysis using *E. coli* expression systems leads on to experimentation with expression in eukaryotic systems. In the particular case of HuEPO, a transient expression system in human embryonic kidney 293 EBNA (HEK 293-EBNA) cell line will provide insights of the potential "portability" of protein surface charge effects in the secretory capacity of a mammalian expression system. Studying the same engineered constructs in different expression systems addresses the question of whether the solubility profile of rHuEPO based on *E. coli* expression is translatable to secretion in mammalian cells.

With the main theme of altering protein solubility using computational approaches developed in our group, the overall aim in this thesis is to provide the first specific experimental tests of these approaches, through protein surface charge patch engineering in relation in two expression systems: *Escherichia coli* and HEK 293-EBNA cells. This will be addressed by four separate papers, which each will encompass a different aim.

Aims:

- To investigate computational and experimental strategies to improve soluble expression of a highly insoluble 6-Phosphofructo-2-Kinase/fructose-2,6-bisphosphatase (PFKFB3) protein through modification of surface features upon heterologous expression in *E. coli*.

❖ This aim has been discussed in paper 1: "Strategies to improve soluble expression of recombinant 6-Phosphofructo-2-Kinase/fructose-2,6-bisphosphatase (PFKFB3) in *E. coli*". In this chapter protein features were explored using structural analysis and computational tools to calculate and modify surface charge. It has also explored the experimental solubility of a set of PFKFB3 WT and variants when expressed in three different *E. coli* systems.

- To analyse the role of positively-charged patches on recombinant human erythropoietin (rHuEPO) surface in experimental protein solubility when expressed in the cytoplasm of *E. coli*.

❖ This aim has been addressed in paper 2: "Increasing solubility in recombinant erythropoietin through modification of surface patches". This paper has examined the role of positively-charged patches on surfaces of rHuEPO WT and variants by applying an algorithm developed by our group. It has also investigated the solubility of rHuEPO WT and variants upon expression in two *E. coli* strains.

- To explore the translatability of the rHuEPO WT and variants aggregation profile in *E. coli* to secretion in mammalian cells.

❖ This aim has been addressed in paper 3: "Modulation of recombinant erythropoietin secretion in HEK 293-EBNA cells through modification of protein surface patches". This paper has investigated the role of positively-charged patches size on rHuEPO surface in the secretory capacity of human embryonic kidney 293 EBNA under transient expression.

- To determine the relative influence of protein surface lysines and arginines on protein solubility.

❖ This aim has been addressed in paper 4: "Alteration of lysine and arginine content as a strategy to modify protein solubility: a test for *E. coli* proteins". This paper has studied the importance of lysines relative to arginines on protein surfaces by diminishing solubility of three *E. coli* proteins: thioredoxin, cold shock-like protein cspB, and the histidine-containing phosphocarrier protein.

1.8 Alternative format

The alternative format thesis has been accomplished in agreement with the rules and regulations of the University of Manchester. The experimental outcomes in this thesis (Chapter 3-6) are being presented as four self-contained papers in the style suitable for the intended journal. To deliver a detailed description of the computational algorithm behind all the results chapters, an extra introductory chapter has been included (Chapter 2). The details of each manuscript are described as follows:

Chapter 3: Strategies to improve soluble expression of recombinant 6-Phosphofructo-2-Kinase/fructose-2,6-bisphosphatase (PFKFB3) in *E. coli*.

Authors: Carballo-Amador M.A., McKenzie E.A., Dickson A.J. and Warwicker J.

Projected journal: Protein Engineering, Design and Selection (PEDS)

Contribution of authors: This manuscript is the outcome of computational and experimental work of which I contributed the majority. This manuscript represent an experimental continuation of a preliminary research exploring PFKFB3 solubility by the second author Dr. Edward McKenzie. Also, this paper is an extension of some initial computational observations published recently from the Warwicker group. However, they did not contribute to any of the computational or experimental design displayed in these chapters with the exception of Dr. Warwicker. My two supervisors Prof. Alan Dickson and Dr. Jim Warwicker provided advice and guidance on all experimental and computational work, respectively. As first author, I was fully responsible for writing the manuscript. Then, the first draft was reviewed and discussed

by my supervisors. These comments were fundamental to develop the final version of this thesis.

Chapter 4: Increasing solubility in recombinant erythropoietin through modification of surface patches.

Authors: Carballo-Amador M.A., Warwicker J. and Dickson A.J.

Projected journal: Protein Engineering, Design and Selection (PEDS)

Contribution of authors: This manuscript is the result of computational and experimental work of which I contributed the majority. My two supervisors Prof. Alan Dickson and Dr. Jim Warwicker provided advice and guidance on all experimental and computational work, respectively. This manuscript represents an experimental continuation of a computational algorithm (charged patches modification) developed recently from the Warwicker group which is addressed in Chapter 3. As first author, I was fully responsible for writing the manuscript. Then, the first draft was reviewed and discussed by my supervisors. These comments were fundamental to develop the final version of this thesis.

Chapter 5: Modulation of recombinant erythropoietin secretion in HEK 293-EBNA cells through modification of protein surface patches.

Authors: Carballo-Amador M.A., Warwicker J. and Dickson A.J.

Projected journal: Protein Engineering, Design and Selection (PEDS)

Contribution of authors: This manuscript is descriptive of experiments of which I contributed the majority. My two supervisors Prof. Alan Dickson and Dr. Jim Warwicker provided advice

and guidance on all experimental and computational work, respectively. This manuscript represents an experimental continuation of some initial computational calculations established in Chapter 4. As first author on this paper, I was fully responsible for writing the manuscript. Then, the first draft was reviewed and discussed by my supervisors. These comments were fundamental to develop the final version of this thesis.

Chapter 6: Alteration of lysine and arginine content as a strategy to modify protein solubility: a test for *E. coli* proteins.

Authors: Carballo-Amador M.A., Dickson A.J. and Warwicker J.

Projected journal: Protein Engineering, Design and Selection (PEDS)

Contribution of authors: This manuscript is the outcome of computational and experimental work of which I contributed the majority. My two supervisors Prof. Alan Dickson and Dr. Jim Warwicker provided advice and guidance on all experimental and computational work, respectively. This manuscript represents an experimental continuation of an initial computational observation published recently from the Warwicker group. However, they did not contribute to any of the computational or experimental design displayed in this chapter with the exception of Dr. Warwicker. As first author on this paper, I was fully responsible for writing the manuscript. Then, the first draft was reviewed and discussed by my supervisors. These comments were fundamental to develop the final version of this thesis.

Chapter 2

Computational approach underpinning the research presented in the experimental papers

Protein design is a consequence of lessons in molecular engineering learned from nature. As more extensive and intensive studies are made, more concepts in synthetic biology will arise. This accumulation of knowledge has been steadily growing in parallel with advances in technology. Over the past decades several computational approaches has been developed, which focused on the analysis of biophysical properties in proteins and the potential relationship to structure and function. Some of these computational studies include important advances in protein structure modelling, stability, folding, aggregation, solubility, post-translational modifications, and function (Blom et al., 2004, Hwang and Park, 2008). These approaches have significant implications for various aspects of recombinant protein technology, whether in pure research or industry. In addition the wide study of expression systems and experimental databases has brought improvements to the prediction tools and this has been enhanced by constant feedback between experimental and computational studies in order to improve the expression and production of proteins.

In this thesis we have applied two recently published hypotheses from our group. One stated that soluble expression of proteins was inversely correlated with the size of the largest positively-charged patch on the protein surface (Chan et al., 2013). The second hypothesis (of protein solubility), arose from the finding that the relative content of lysine and arginine residues separated the *E. coli* proteins by solubility (Warwicker et al., 2014). Both hypotheses arose from a study of an extensive dataset of experimental solubilities determined for cell-free

expression of *E. coli* proteins (Niwa et al., 2009). In combination with other widely used strategies, such as lowering expression temperature and inducer concentration, decreasing non-charged (hydrophobic) patches and addition of helical capping for increasing stability, we have explored a rational understanding for directed alteration of solubility in a variety of recombinant proteins.

2.1 Computational tools to develop stable recombinant proteins

Despite the breakthroughs in genetic (Kwaks and Otte, 2006) and cellular engineering (Tigges and Fussenegger, 2006), there still remains a need to develop proteins with improved activity and stability (Barnes and Dickson, 2006, Tokuriki et al., 2008). Protein instability is closely related to protein aggregation, since this may arise from partially folded intermediates as a consequence of the destabilisation of the native conformation (Fink, 1998). One major concern around recombinant protein manufacture is protein aggregation (low solubility), since this can lead to decreased protein expression, diminish biological activity and increase immunogenicity (Wu et al., 2010). A novel area is emerging to predict and solve these issues, and informatics software is being introduced to predict physicochemical properties in proteins, such as glycosylation and phosphorylation sites (Blom et al., 2004), protein solubility (Smialowski et al., 2012, Chan et al., 2013, Chang et al., 2014), aggregation-prone regions (Voynov et al., 2009, Weiss et al., 2009), and protein folding-misfolding (Schymkowitz et al., 2005).

Computational tools have been progressing to develop more accurate predictions in protein design (Hwang and Park, 2008). Protein design, whether based around structure or function, starts with a correlation analysis between protein structure, function, stability, and sequence (Park et al., 2004). Computational design and simulation of recombinant proteins has

become a valuable tool for evaluating experimental data and developing models (Hansmann, 2008). Some of the improvements in therapeutics cover antibody affinity, protein-protein interaction, incrementing stability and diminishing protein aggregation (Table 2.1) (Hwang and Park, 2008). In the particular case of antibodies, it is preferred to have a crystal structure, or a model, of the antibody to build an accurate computational method (Caravella et al., 2010). Voynov and co-workers developed a method called "spatial aggregation propensity (SAP)" that calculated the dynamic exposure of hydrophobic patches (Voynov et al., 2009). This method identifies hot-spots for aggregation, based on clusters of hydrophobic amino acids. The same group showed that, using SAP prediction, they can perform target mutations in those regions to enhance antibody stability (Chennamsetty et al., 2009). The utility of the SAP technology has been applied to design cysteine variants in antibodies (Voynov et al., 2010). These applications at the antibody surface are favourable due to the presence of cysteine residues, which are involved in disulphide bridges between light and heavy chains. SAP has been used to predict protein binding regions in immunoglobulin (IgG1) and epidermal growth factor receptor (EGFR) (Chennamsetty et al., 2011). In addition to SAP, a series of algorithms have been developed to potentially address or understand protein aggregation issues, such as AGRRESCAN (Conchillo-Sole et al., 2007), PASTA and PASTA 2.0 server (Trovato et al., 2007, Walsh et al., 2014), TANGO (Fernandez-Escamilla et al., 2004) and Zyggregator (Tartaglia and Vendruscolo, 2008).

In the affinity enhancements area, Lippow et al. demonstrated that an antibody (cetuximab) was 10-fold more potent than its wild-type by targeting mutation of the complementarity determining region (Lippow et al., 2007). In addition, other successful experimental or computational design approaches have been applied to improve the affinity of antibodies (Table 2.1) (Clark et al., 2006, Barderas et al., 2008, Farady et al., 2009). In order to obtain a better understanding of protein structure-function or to locate potential amino acids

to mutate, conservation assays among homologous proteins have been useful (Roosild et al., 2006). These assays have been used to identify critical residues for protein folding and structure (Asano et al., 2011).

Table 2.1. Summary of structural modifications in therapeutic proteins.

Recombinant therapeutic protein	Computational Method	Structural Modification	Molecular Improvement	References
Monoclonal Antibodies	Interactive computational design	Target mutations of CDR	Affinity	(Clark et al., 2006)
	Interactive computational design	Target mutations of CDR	Affinity	(Lippow et al., 2007)
	Structural modelling; semi-automated tool for antibody construction	Target mutations of CDR	Affinity	(Barderas et al., 2008)
	Structured base and algorithms design	Target mutations of DCR	Affinity	(Farady et al., 2009)
	SAP	Target mutations in hydrophobic regions.	Stability	(Chennamsetty et al., 2009)
	n/a	Targeted mutation of CDR; glycosylation site insertion.	Solubility	(Pepinsky et al., 2010)
	n/a	pI modification; hydrophobic surfaces reduction; glycosylation site insertion.	Solubility	(Wu et al., 2010)
EPO	n/a	Addition of 1-4 glycosylation sites	Half-life and B. activity	(Su et al., 2010)
	n/a	Removed 3 *N*-glycosylation sites	Decrease aggregation in *E. coli*	(Narhi et al., 2001)
	SAP	Four mutations using random mutagenesis and ribosome display	B. activity and less prone to aggregate	(Buchanan et al., 2012)
Factor VIII	n/a	Three aa substitution	B. activity	(Allen et al., 2007)
	n/a	Combining mutations	Stability	(Wakabayashi et al., 2009)
Insulin	n/a	Specific aa; changing charge or hydrophobicity	Rapidly absorbance	(Hirsch, 2005, Sheldon et al., 2009)
tPA	n/a	Binding site aa replacement	Half-life	(Dunn and Goa, 2001)

Note. aa, amino acids; B. activity, Biological activity; CDR, Complementarity determining region; pI, isoelectric point; SAP, spatial aggregation propensity; tPA, tissue plasminogen activator.

2.2 Protein solubility predictors

The first method to calculate protein solubility from sequence was proposed by Wilkinson and Harris (Wilkinson and Harrison, 1991). This model is based on average charge (determined by Asp, Glu, Lys and Arg residues) and the content of turn-forming residues (Asn, Gly, Pro and Ser). In the past years, a number of algorithms to predict protein solubility have been generated based on amino acid sequences (Davis et al., 1999, Smialowski et al., 2007, Magnan et al., 2009). Idicula-Thomas and Balaji developed a code based on protein sequence and structure (Idicula-Thomas and Balaji, 2005). These computational codes are designed to assess the solubility of a protein upon recombinant expression in *Escherichia coli*. In a study performed by Magnan et al., (2009), they reported the prediction accuracy of a revised Wilkinson-Harrison (RevW-H) model showing an accuracy prediction of 53.7%, the PROSO predictor (Smialowski et al., 2007) has shown a maximum accuracy of 59.3%, and SOLpro achieves 74.2% (Magnan et al., 2009). In a recent study Chang et al., (2014) compared the use of a uniform data set to test series of prediction tools, including an updated version PROSO II (Smialowski et al., 2012) and ccSOL (Agostini et al., 2012) predictors. The predictions revealed an accuracy of 51.4% for RevW-H, 54.2% for ccSOL, 57.8% for PROSO, 59.9% for SOLpro, and 64.3 for PROSO II (Chang et al., 2014).

Despite the growth of solubility machine-learning-based models, there is no consensus approach in general protein design to answer the question "which proteins will have a better or worse chance of being expressed in a soluble form in heterologous systems?". Nevertheless, these and further algorithms may achieve more accuracy and reliability through experimental validation. For example, the use of experimental solubility databases, such as those of Taguchi's group (Niwa et al., 2009), have been helpful in building algorithms, e.g., ccSOL (Agostini et al., 2012), and our two surface charge calculation approaches (Chan et al., 2013, Warwicker et al., 2014).

2.3 Experimental database for solubility predictors

The solubility database of all *E. coli* proteins (eSOL) (Niwa et al., 2009) has inspired the development of a series of computational approaches (Agostini et al., 2012, Samak et al., 2012, Chan et al., 2013, Fang and Fang, 2013, Agostini et al., 2014, Klus et al., 2014, Warwicker et al., 2014). The eSOL database came from the study of the complete *E. coli* open reading frame library (ASKA library) (Kitagawa et al., 2006) expression in a cell-free translation system (PURE system) (Shimizu et al., 2005). This is a reconstituted system that only involves essential *E. coli* elements responsible for protein synthesis (Fig. 2.1). They successfully quantified around 70% of the ASKA library, which comprised 3,173 proteins (out of a potential 4,132 proteins). The remaining 30% of the library was not quantified due to insufficient translation material and technical problems. These studies revealed a bimodal distribution comprising two categories, defined by the authors as soluble and aggregation-prone groups. In addition, the same group published a second approach incorporating chaperones in the PURE system (Niwa et al., 2012). This cell-free expression system containing chaperones may be useful to investigate properties (sequence or structural-based) of target proteins in order to understand the chaperone influence in their solubilities.

Our group has used the eSOL database to look at both sequence and structure-based features that best separate the most and least soluble subsets (Chan et al., 2013, Warwicker et al., 2014). These strategies offer synthetic redesign using the predictive information generated by both algorithms.

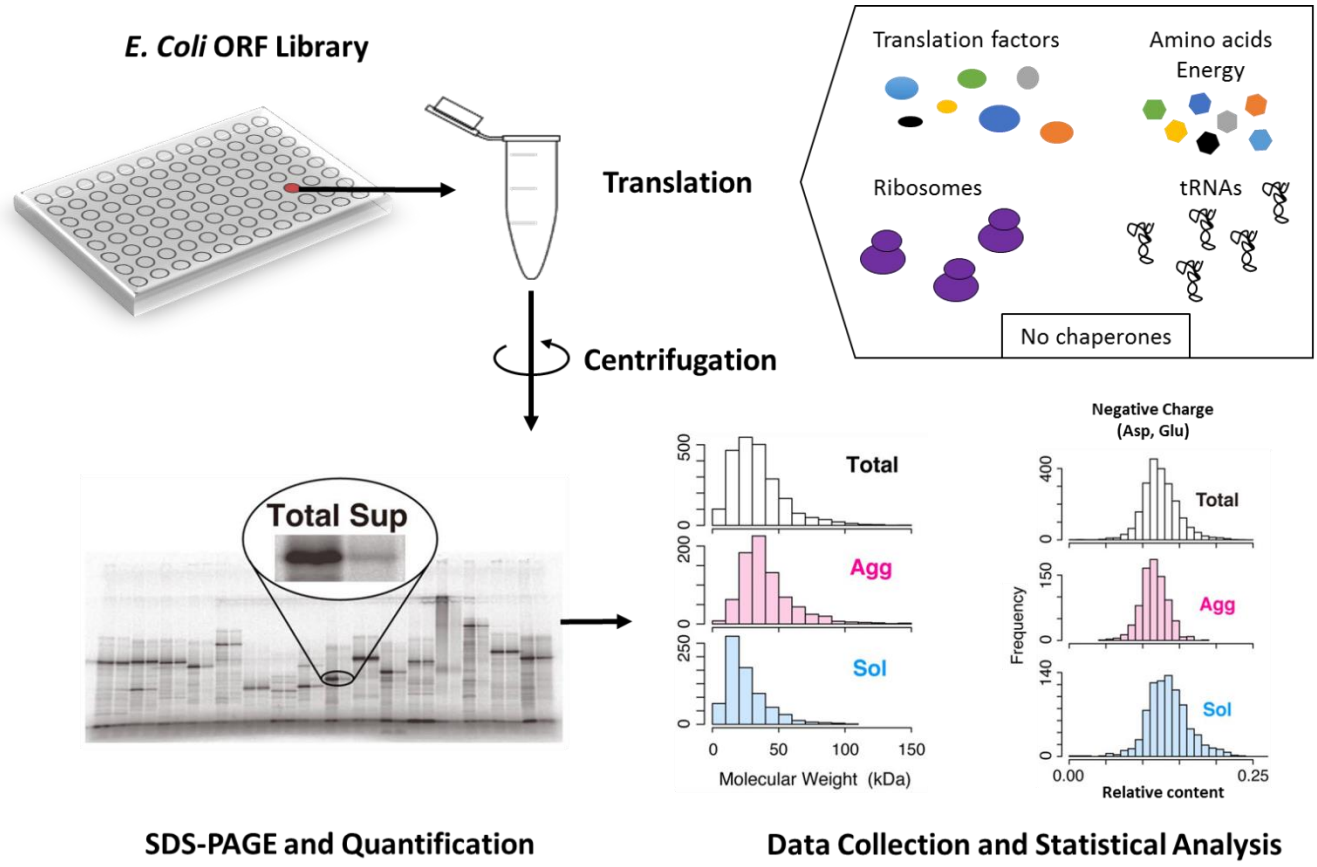


Fig. 2.1. Schematic representation of the *in vitro* proteome expression for aggregation analysis. Each ORF from the ASKA library was expressed in the reconstituted cell-free system (PURE system) for 60 minutes. The translation products were labelled with [^{35}S]methionine. After translation, samples were centrifuged to obtain the soluble fraction (i.e. supernatant fraction). The uncentrifuged fraction (Total) and supernatant (Sup) were subjected to separation by SDS-PAGE. The protein bands were quantified by autoradiography. Data collection and statistical analysis show a total solubility relative to the molecular weight histogram and the relative contents of negatively-charged residues (Asp and Glu) for the 3,173 quantified proteins. Protein solubility was defined as the intensity ratio in the Total and Sup fractions. Those proteins with solubilities <30% and >70% were classified as the aggregation-prone (Agg, coloured pink) and soluble (Sol, coloured blue) sets, respectively (Adapted from Niwa et al., 2009 and Niwa et al., 2012).

2.4 Surface charge calculations and protein solubility

2.4.1 A correlation between positively-charged patches and insolubility

A published algorithm from our group (Chan et al., 2013) focuses on sequence and structure-based properties analysis from the experimental eSOL database (Niwa et al., 2009). The key feature in this computational approach is the calculation of surface charges, which include the nonQmax –the maximal size of a non-charged patch-, posQmax –the maximal positively charged patch- and their multiplicative combination, versus thresholds calculated from the eSOL protein dataset (Niwa et al., 2009). These thresholds have been derived from the value of parameters that best separate less and more soluble proteins. By far the best separation for an individual feature was that with largest positively-charged patch, for which proteins with the largest positive patch predicted as least soluble (ratio above 1.0 relative to the threshold) (Fig. 2.2). Positively-charged patches with ratio below 1.0 are predicted as more soluble. The outcome prediction is a full mapping of the electrostatic potential patches on the protein surface (Fig. 2.2). This allows the adjustment of protein solubility by altering posQ with positively- or negatively-charged amino acid mutations. This section relates to the work presented in the experimental Chapters 3, 4 and 5.

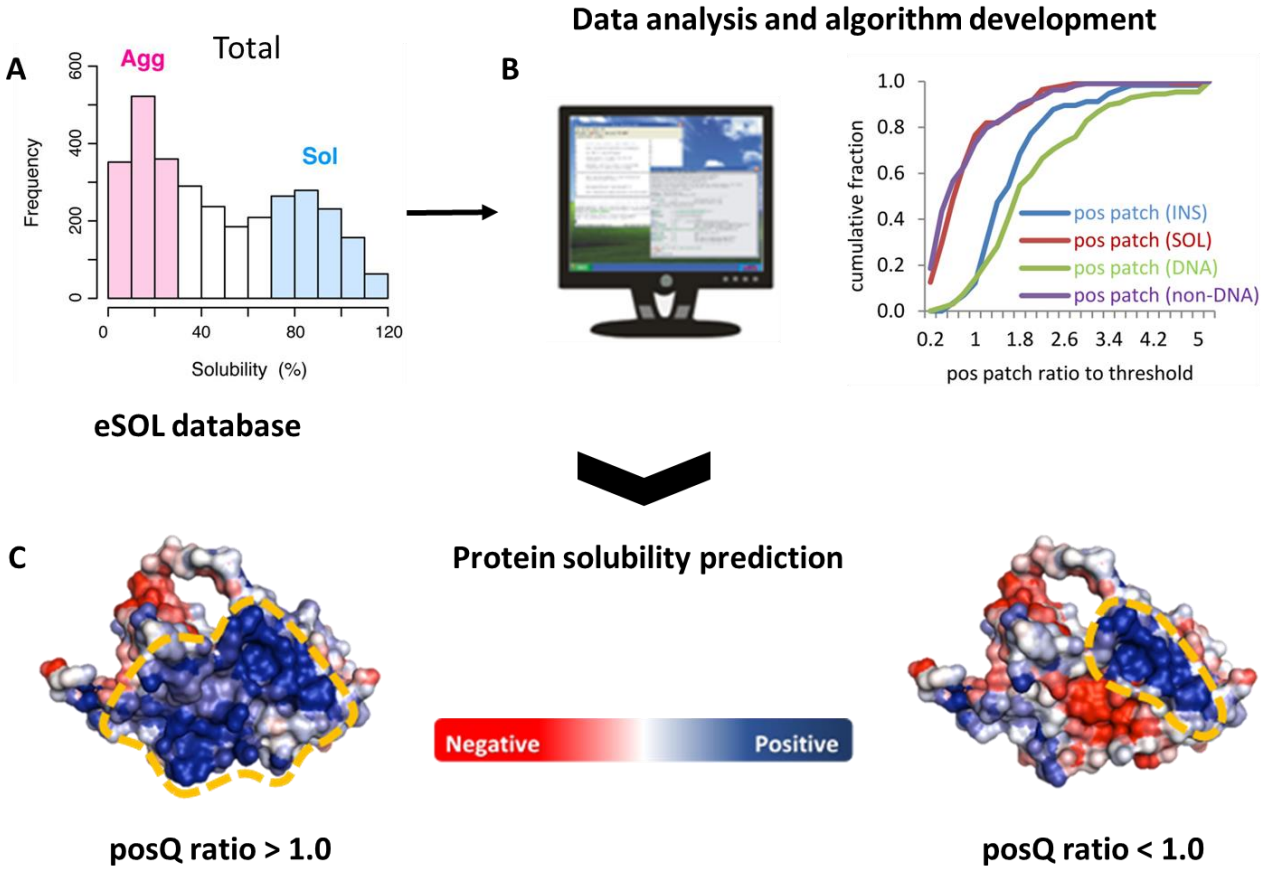


Fig. 2.2. Overview of the computational approach development and application. (**A**) Histogram of solubility distribution for 3,173 quantified proteins from the eSOL database. (**B**) Data collection and statistical analysis showing the positive patch distribution to threshold, alongside with DNA-binding and non DNA-binding calculations. (**C**) Protein surface structures showing the largest positively-charged patches surface distribution, insoluble (posQ >1.0) and soluble (posQ <1.0) prediction on the left and right, respectively. Amino acids in the largest positive patches are represented by blue, non-charged patches by white and negatively charged by red colour, respectively. Adapted from Niwa et al., 2009 and Chan et al., 2013.

2.4.2 Sequence-based property of lysine versus arginine content separated the *E. coli* protein least and most soluble subsets

A second computational approach arose from the previous correlation from largest positively-charged patches and insolubility analysis, observing that those basic patches are enriched for arginine, compare to lysine (Chan et al., 2013). From this observation, our group found that the sequence-based property of lysine *versus* arginine content also separated the *E. coli* protein into their least and more soluble subsets (Fig. 2.3) (Warwicker et al., 2014). This resulted in the hypothesis stating that lysines are more favourable for protein solubility than are arginines. In addition, this statement correlates with the observation that proteins occurring at higher concentrations in nature (*versus* those at lower concentrations) also had a relative preference for lysine (Warwicker et al., 2014). This lead to the suggestion that the balance of lysine and arginine could be a general effect, not just specific to expression systems. In support of this, we noted a study in which the change of lysines to arginines in GFP resulted in decreased solubility when the arginine-rich GFP was expressed in *E. coli* (Sokalingam et al., 2012). This section relates to the study presented in the experimental Chapters 6.

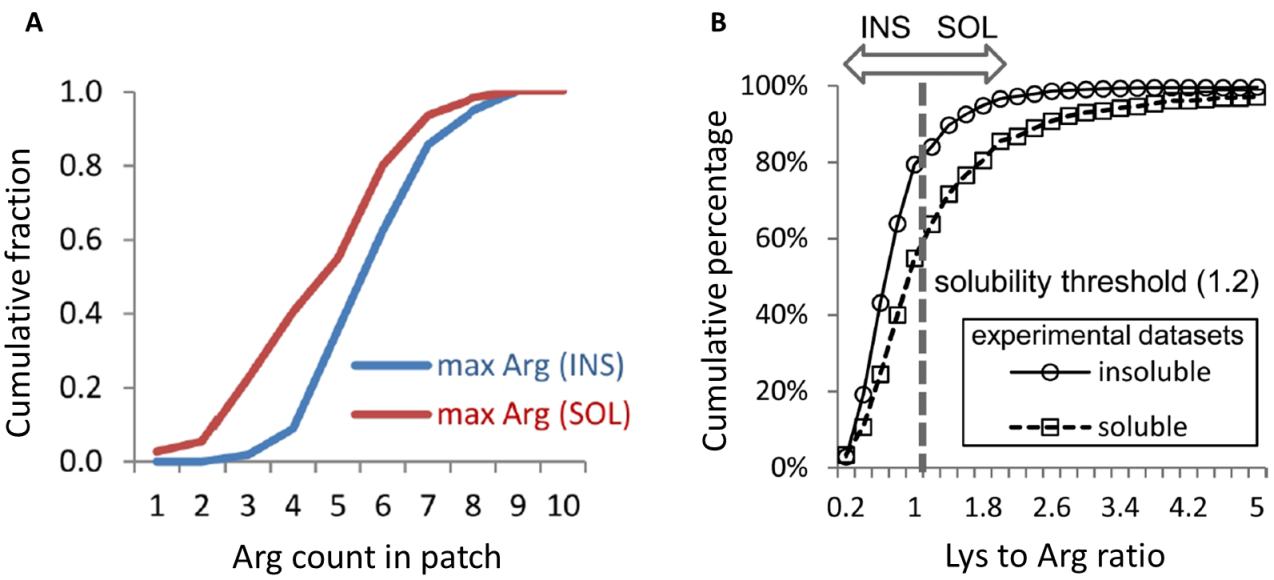


Fig. 2.3. Sequence-based analysis of the separation of lysine to arginine ratio for soluble and insoluble distribution from eSOL database. (**A**) Largest arginine content separation according to the charged patch size. (**B**) Separation of KR-ratio for soluble and insoluble, relative to the threshold value (1.2) (Adapted from Chan et al., 2013 and Warwicker et al., 2014).

In the following chapters, the experimental application of the two computational observations have been studied. The first approach, altering positively-charged patches on protein surface, has been employed in all the Chapters, except for the final chapter (Chapter 6). The second approach, altering the lysine to arginine ratio, is the backbone of Chapter 6, in a study of *E. coli* protein solubilities. Both computational applications constitute the basis for the protein design behind the experimental Chapters in this thesis.

Chapter 3

Paper 1:

Strategies to improve soluble expression of recombinant 6-Phosphofructo-2-Kinase/fructose-2,6-bisphosphatase (PFKFB3) in *E. coli*

Carballo-Amador M.A.[12], McKenzie E.A.[2], Dickson A.J.[1] and Warwicker J.[2]

[1] Faculty of Life Sciences, University of Manchester, Michael Smith Building, Oxford Road, Manchester M13 9PT, UK

[2] Faculty of Life Sciences, University of Manchester, Manchester Institute of Biotechnology, 131 Princess Street, Manchester M1 7DN, UK

Abstract

Production of highly concentrated and soluble proteins is a necessary requirement whether for biophysical and structural studies or for therapeutic purposes. *Escherichia coli* offers a cost-efficient expression system, however, it is unable to carry out post-translational modifications such as occurs in eukaryotic cells. One main limitation is the formation of insoluble protein aggregates (inclusion bodies) in the cytoplasm. Based on protein structure, an algorithm was developed in our group to predict protein solubility, calculating polar and non-polar patches on the protein surface. Employing this algorithm, we predicted amino acid mutations that would facilitate expression of variants of recombinant 6-Phosphofructo-2-Kinase/fructose-2,6-bisphosphatase (rPFKFB3) upon heterologous expression in three *E. coli* strains (BL21 (DE3) pLysS, BL21-CodonPlus and SHuffle). In addition, B-factor analysis was performed to introduce a charged helical cap into thermo-flexible areas. Solubility calculations suggested that rPFKFB3 is a highly insoluble protein upon expression in *E. coli*. We found small variations between wild-type and mutant solubilities among the three experimental strategies trialled: (i) a predicted increase polarity, (ii) predicted diminished positively-charged patches and (iii) predicted diminished local flexibility. It does appear, however, that charge properties (positively-charged patches) should be considered for further applications, since mutants with reduced positive patches gave consistently increased solubility.

3.1 Introduction

Four distinct PFKFB genes and isoforms (liver [PFKFB1], heart [PFKFB2], inducible form [PFKFB3] and testis [PFKFB4]) have been characterised in mammalian cells (Hue and Rousseau, 1993, Chesney et al., 1999). For all PFKFB isoforms, two separate domains in the protein modulate the cellular concentration of fructose-2,6-bisphosphate (Fru-2,6-P_2). One domain catalyses the synthesis of Fru-2,6-P_2, using fructose-6-phosphate (F6P) and ATP as substrates at the N-terminal region (2-Kase) and the second C-terminal domain (2-Pase) catalyses the hydrolysis of Fru-2,6-P_2 into F6P and inorganic phosphate (Pilkis et al., 1984, Bazan et al., 1989). The origin of such a "two-in-one" protein species has been speculated to have arisen from the fusion of separate kinase and phosphohydrolase/mutase ancestral genes (Bazan et al., 1989). There is a high sequence identity amongst PFKFB isoenzymes, and they share a well conserved (85%) core activity domain (2-Kase/2-Pase) (El-Maghrabi et al., 2001, Kim et al., 2006). Multiple sequence alignments have been analysed to understand similarities amongst isoforms and ancestral catalytic units. These studies have revealed a homology between PFKFB and phosphohydrolase/mutase at the C-terminal bisphosphatase activity domain (Bazan et al., 1989). Recent studies of this domain in PFKFB3 (residues 440-446) have identified an important role in the binding of substrate and release of the product from the catalytic pocket (Cavalier et al., 2012). Also, the N-terminal unit shares sequence and structure similarities to some nucleotide binding proteins (Hasemann et al., 1996).

Expression and purification of proteins forms a key aspect of detailed structural and functional studies of proteins. Lin et al.(1990) reported that rat hepatic PFKFB1 aggregated in inclusion bodies when expressed in *E. coli* (Lin et al., 1990). Hydrophobic interactions between chemical groups in proteins of various degrees of unfolding or conformational alteration peptides can lead to diminished solubility and aggregation to form multimeric combinations of

high molecular weight (Fink, 1998). The formation of aggregates during heterologous expression can decrease protein yield or generate the need to develop challenging re-solubilisation protocols but also provides indications of proteins that may undergo aggregate-forming molecular interactions during subsequent purification and solution storage (Khan et al., 2013). The propensity for aggregate formation *in vivo* or *in vitro* is determined by multiple factors such as protein amino acid sequence (e.g. influencing electrostatic interactions, hydrophobic [non-charged] patch interactions), pH, temperature, and excipients in solution (e.g. proteins, small molecular weight additives, metal ions and salt concentration) (Fink, 1998, Majhi et al., 2006, Greaves and Warwicker, 2007).

Protein aggregation is considered a potential problem at all stages of protein production and long-term storage (Bondos and Bicknell, 2003, Fowler et al., 2005, Weiss et al., 2009). To achieve a native structure of the protein of interest, a considerable amount of protein (5 – 50 mg) is required in a soluble, monomeric form (Peti and Page, 2007). The literature illustrates several examples of proteins that exhibited limited solubility when expressed in *E. coli* with the resultant unsuccessful crystallisation and failed structural analysis [e.g. HIV-1 integrase (Dyda et al., 1994), human erythropoietin (Cheetham et al., 1998), mitogen activated protein kinase (Patel et al., 2004) and human apolipoprotein D (Nasreen et al., 2006)]. It is illuminating that single or multiple mutations undertaken with each of these proteins generated protein variants of increased solubility that were successfully used in structural analysis. This highlights the potential predictability between the nature of amino acid composition and sequence for success of purification and subsequent structural analysis and also indicates features that may need to taken into account in the rationalisation of design and domain organisation of novel protein formats. Our group has recently published an algorithm (Chan et al., 2013), which comprises structure- and sequence-based feature analysis of an *E. coli* protein solubility database (Niwa et al., 2009). The algorithm computes surface charges (non-charged

[hydrophobic] patch, positively charged patch and their multiplicative combination) versus thresholds calculated from Niwa dataset of experimental solubilities determined for cell-free expression of *E. coli* proteins (Niwa et al., 2009). These thresholds have been derived by looking for the value of a parameter that best separates less and more soluble proteins, with those proteins with the largest positive patch predicted as least soluble.

The Human inducible form of 6-Phosphofructo-2-kinase/fructose-2,6-bisphosphatase 3 (PFKFB3, formerly reported as iPFK2; EC 2.7.1.105 and EC 3.1.3.46), is a 59.6 kDa bifunctional enzyme of central importance in controlling glucose utilisation during cancer cell proliferation (Chesney et al., 1999). PFKFB3 has an extremely high 2-Kase-to-2-Pase activity, in a ratio slightly above 700:1 (Sakakibara et al., 1997) and, due to these properties, PFKFB3 has been highlighted as a potential target for cancer therapy (Kim et al., 2006, Clem et al., 2013). In order to understand the structure-activity relationship in detail, human PFKFB3 has been expressed and purified from *E. coli* BL21 strains (Kim et al., 2006, Clem et al., 2008, Brooke et al., 2014, Yamamoto et al., 2014). These reports lacked information about protein solubility, or of the propensity of the protein to aggregate in the cytosol of *E. coli*. Previous experimental attempts to express PFKFB3 in a soluble form in *E. coli* have been problematic (McKenzie E.A., unpublished data). Based on this, a set of proteins were tested using the algorithm developed in our group (Chan et al., 2013). From these proteins, which experimental solubility profile was known for *E. coli* expression (comprising bad to good soluble recombinant proteins) (McKenzie E.A., unpublished data), PFKFB3 was the most insoluble protein, agreeing with the experimental work.

In this manuscript, we describe three different strategies to improve the expression of the human inducible form of PFKFB3 expression in *E. coli*. The strategies involve synthetic redesign of the amino acid composition utilising the predictive information generated by use of the algorithm developed in our group (Chan et al., 2013) to (i) decrease non-polar areas, (ii)

diminish positively-charged patches and (iii) introduce a charged helical cap into thermo-flexible areas. PFKFB3 was chosen as presenting a challenging, therapeutically-valid target prone to aggregate peptide from both our experimental studies and preliminary computational predictions. The aim was to analyse the protein structure and surface as starting point for a rational design of improved mutants. As far as we know, this is the first study of any PFKFB isoform based on predicted calculation to enhance protein solubility and stability.

3.2 Materials and methods

3.2.1 *PFKFB3 structural analysis and solubility prediction*

All structural analysis were carried out by using the open source Swiss-PdbViewer 4.0.1 (Guex and Peitsch, 1997) and PyMOL Molecular Graphics System educational version 1.3 (Schrödinger, 2010) from PFKFB crystal structures (PDB ID: 2AXN and 1K6M). Aggregation-prone patches (largest non-charge patch, nonQ) and protein solubility profile (largest positively-charged patch, posQ) (Supplementary Table S3.1) were predicted using software tools developed in our group (Chan et al., 2013). In addition to PFKFB3 (PDB ID: 2AXN), a structure (PDB ID: 1K6M) of the liver isoform PFKFB1 was also used in analysis of B-factors. It was rationalised that crystal contacts may vary between the two structures, and looking at an additional structure and its flexibility as reflected in B-factors, could be valuable. Multiple sequence alignments were performed using ConSurf server (Ashkenazy et al., 2010, Celniker et al., 2013).

3.2.2 *Protein engineering and expression vectors construction*

A modified version of the commercial pET-16b vector (Novagen) pHis vector encoding a thrombin-cleavable amino-terminal 6x His tag (Section 9.2) was used to express wild-type and mutated forms of PFKFB3. PFKFB3 mutations were carried out using GENEART Site-Directed Mutagenesis System with the enzyme AccuPrime *Pfx* (Invitrogen). The resulting mutants were as follows: rPFKFB3 M1 (W13Y, V14N, V16K), rPFKFB3 M2 (Q100E, L103E, A104E), rPFKFB3 M3 (N91D, E92Q, N178D) and rPFKFB3 M4 (R427D). The cloned gene sequence for each plasmid was as follows: 5'-6xHis-Thrombin cleavage site-*BamHI*-PFKFB3-*EcoRI*-3'. Every single base substitution was done by the rationale selection of the GenScript human codon usage frequence table in order to maintain the optimal human gene.

3.2.3 *Protein expression and solubility assay*

All proteins were expressed in three *E. coli* strains: BL21 (DE3) pLysS, BL21-CodonPlus and SHuffle. Bacteria were grown in 50 ml conical tubes containing 5 ml of Luria-Bertani broth (LB) supplemented with 100µg/ml ampicillin with or without 50µg/ml chloramphenicol, as appropriate for plasmid maintenance. After growing for 14-16 h at 37°C, 2% (v/v) of the pre-culture were transfer to a 250 ml conical flask containing 50 ml of LB medium supplemented with 2% (w/v) glucose and 100µg/ml of ampicillin. Cells were grown to an OD_{600} of 0.6 to 0.8 at constant temperature of 25°C. Protein expression was induced with 0.05 mM IPTG. After growing for 5 h post-induction, cells were harvested by centrifugation at 6,500 g for 15 min at 4°C. Bacterial pellets were suspended in 5 ml of lysis buffer (25 mM Tris pH 7.5, 150 mM NaCl, 1% Triton X-100) and were stored at -20ºC until further use. The cells were subjected to sonication for cell disruption by eight cycles of 30 s at 20% amplitude and then allowed to cool

for 30 s on ice water bath. The resulting suspension was centrifuged at 18,000 g for 30 min at 4°C. The supernatants were collected and handled as the soluble fraction. Uncentrifuged samples were handled as the total fraction.

3.2.4 *Western blotting*

Protein in soluble and total fractions was analysed by sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE) in a 12% (w/v) separating gel with a 5% (w/v) stacking gel using the Mini-PROTEAN Tetra Cell (BioRad). Protein samples were mixed with 6x loading buffer (375 mM Tris pH 6.8, 12% [w/v] SDS, 60% [v/v] glycerol, 0.06% [w/v] bromophenol blue) while for reduced samples β-mercaptoethanol was added each time to a final concentration of 5.5% (v/v), followed by incubation at 95°C by 5 min. Equal volumes of total protein and soluble fractions were separated in electrode running buffer (50 mM Tris, 0.38 M glycine, 0.2% (w/v) SDS) at 60 V until samples passed into the separating gel and then the voltage was increased to 160 V at room temperature. For the specific detection of PFKFB3, proteins were transferred to nitrocellulose membrane surrounded by transfer pads (BIO-RAD) that were soaked into blotting buffer (25 mM Tris, pH7.4, 0.2 M glycine and 20% (v/v) methanol) after their separation by SDS-PAGE. The transfer was performed using a transblot semi-dry transfer cell (Bio-Rad) at 15V for 45 min. After blocking the membrane in blocking buffer (5% [w/v] skimmed milk in TBS-Tween pH 7.4) for 12-14 h at 4°C with shaking, the membrane was incubated for 2 h in blocking buffer solution containing mouse anti-polyHis (Sigma) (1:5000 dilution) at room temperature in agitation. For detection of PFKFB3, an IR-labelled secondary Donkey anti-Mouse IgG antibody (LI-COR) (1:15000 dilution) in blocking buffer solution was added for 45 min after the removal of the primary antibody by washing the

membrane with TBS-Tween. For IR detection, blots were imaged with the Odyssey Imaging System. Bands were quantified in the Image Studio Lite software (LI-COR) in order to estimate protein solubility and relative total expression.

3.3 Results

3.3.1 *The effect of decreasing the largest non-polar patch in PFKFB3*

Decreasing non-polar (hydrophobic) areas on protein surface is a well-known practice to increase solubility. The area highlighted in Fig. 3.1 is the most non-polar area of the PFKFB3's surface, and scores highly in the non-polar to polar ratio based on the patch calculations (Chan et al., 2013). This area at the N-terminal extension includes residues L2, E3, L4, W13, P15 and V16. This is a good target region for mutagenesis in order to improve polarity since is not entirely conserved (Fig. 3.2) and away from the active domains (Supplementary Fig. S3.1). After a series of structural analysis and non-polar to polar ratio calculations, substitutions W13Y, V14N and V16K showed an improvement in polarity and without identifiable compromise to stability and judged from the dimer interface in the 2AXN crystal structure (Kim et al., 2006). We applied these mutations on surface to obtain a mutant with improved solubility (i.e. less aggregation-prone) when expressed in *E. coli* (Fig. 3.1B). PFKFB3 wild-type and M1 were transformed into *E. coli* BL21 (DE3) pLysS, BL21-CodonPlus and SHuffle strains. Inclusion body (IB) formation were found indirectly by the difference between total and soluble fraction (Fig. 3.3A-C). Based on this, IB are more present in SHuffle and BL21-CodonPlus (Fig. 3.3D). On the other hand, even though in BL21 (DE3) pLysS is relative more

soluble, it is still insoluble according to the background of our algorithm, based on Niwa et al. classification (soluble >70%) (Niwa et al., 2009). A significant decrease in protein solubility was observed for the M1 in BL21 (DE3) pLysS (Fig. 3.3D).

As presented in Fig. 3.3E, there is a constant but not significant increase in relative total rPFKFB3 amount for wild-type and M1 when expressed in SHuffle and Bl21-CodonPlus strains compared to BL21 (DE3) pLysS (Table 3.1). This relative total rPFKFB3 production was quantified by densitometric analysis and plotted as arbitrary units.

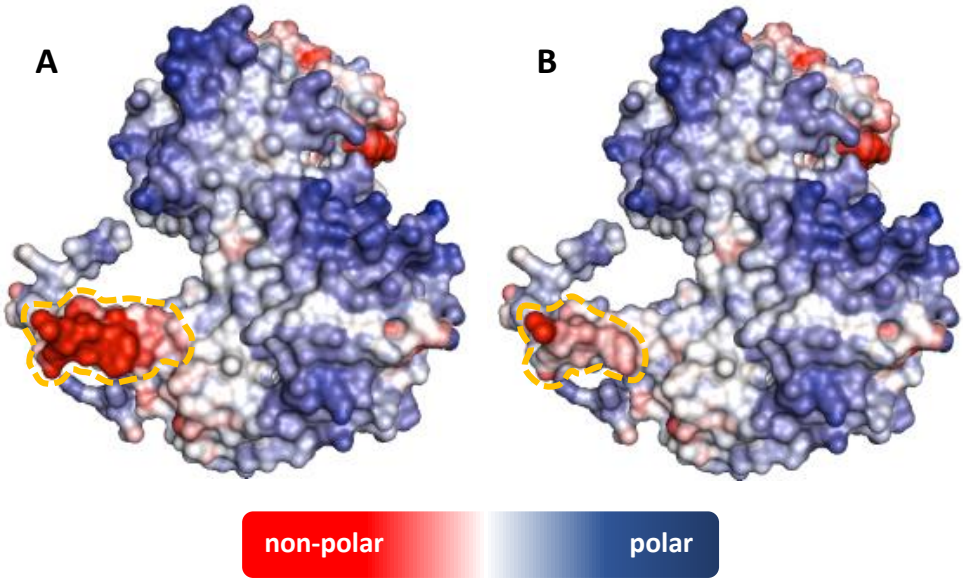


Fig. 3.1. Diminishing non-polar patches on the native protein surface (PDB ID: 2AXN). Structures of the molecular surface of (**A**) PFKFB3 wild-type and (**B**) enhanced variant M1, demonstrating the breakage of the non-polar region around mutated sites at the amino-terminal. Orange dashes lines indicate the patches calculated from the algorithm.

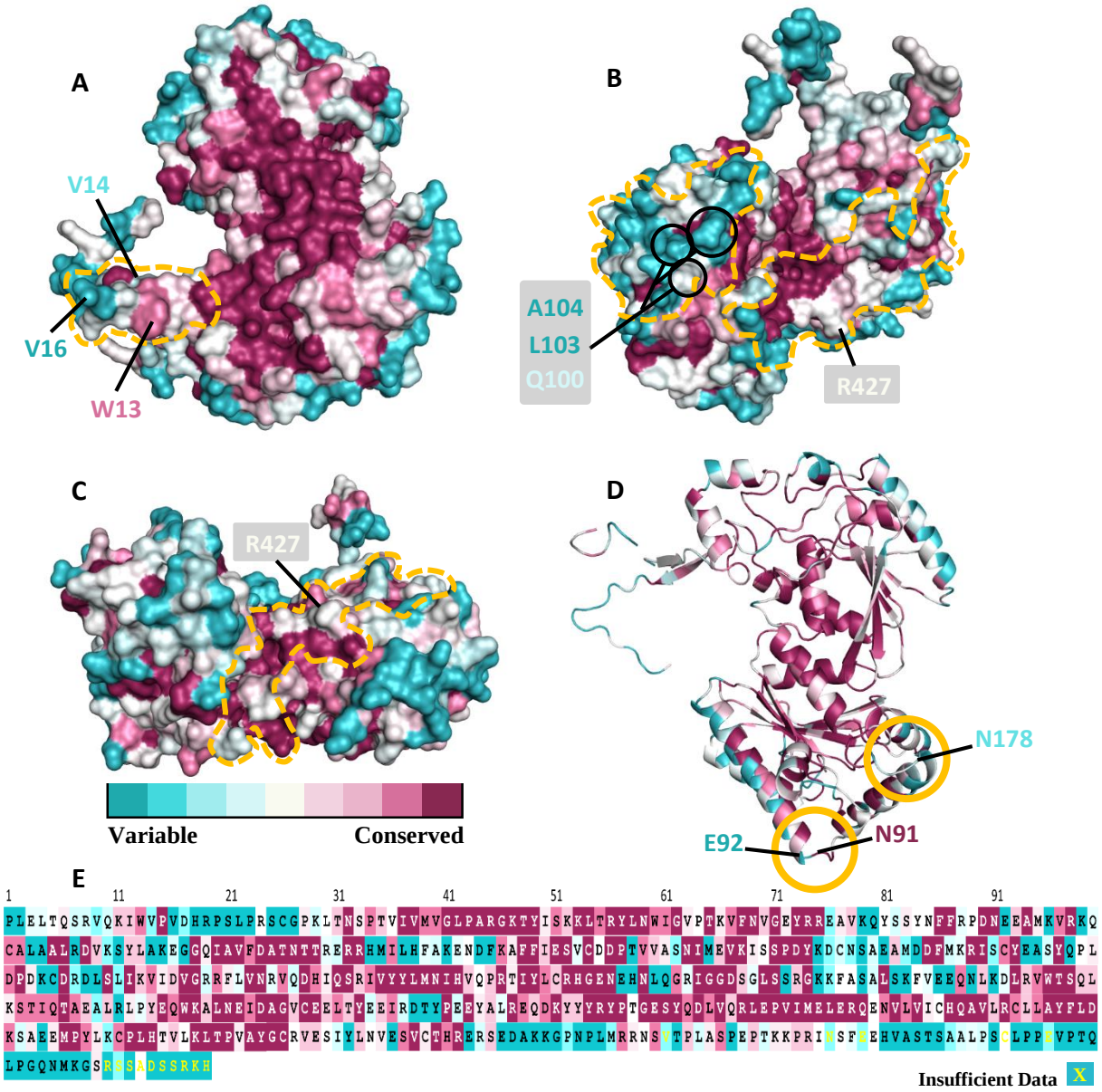


Fig. 3.2. Structure and sequence map coloured by residue conservation scores of PFKFB3 wild-type. (**A**) Conservation visualisation of amino acids in the most non-polar surface patch. This area contains variable residues (V14 and V16) and a non-highly conserved tryptophan (W13). (**B-C**) Residues on the positively-charged patches are less conserved, demonstrating a variability in the four selected targets (Q100, L103, A104 and R427). (**D**) Localisation of the most thermal-flexible areas (orange circles), showing variable helical capping residues (E92 and N178). (**E**) Sequence conservation of PFKFB3 wild-type. Conservation was calculated using ConSurf server (Ashkenazy et al., 2010). Orange dashes lines indicate the patches calculated from the algorithm.

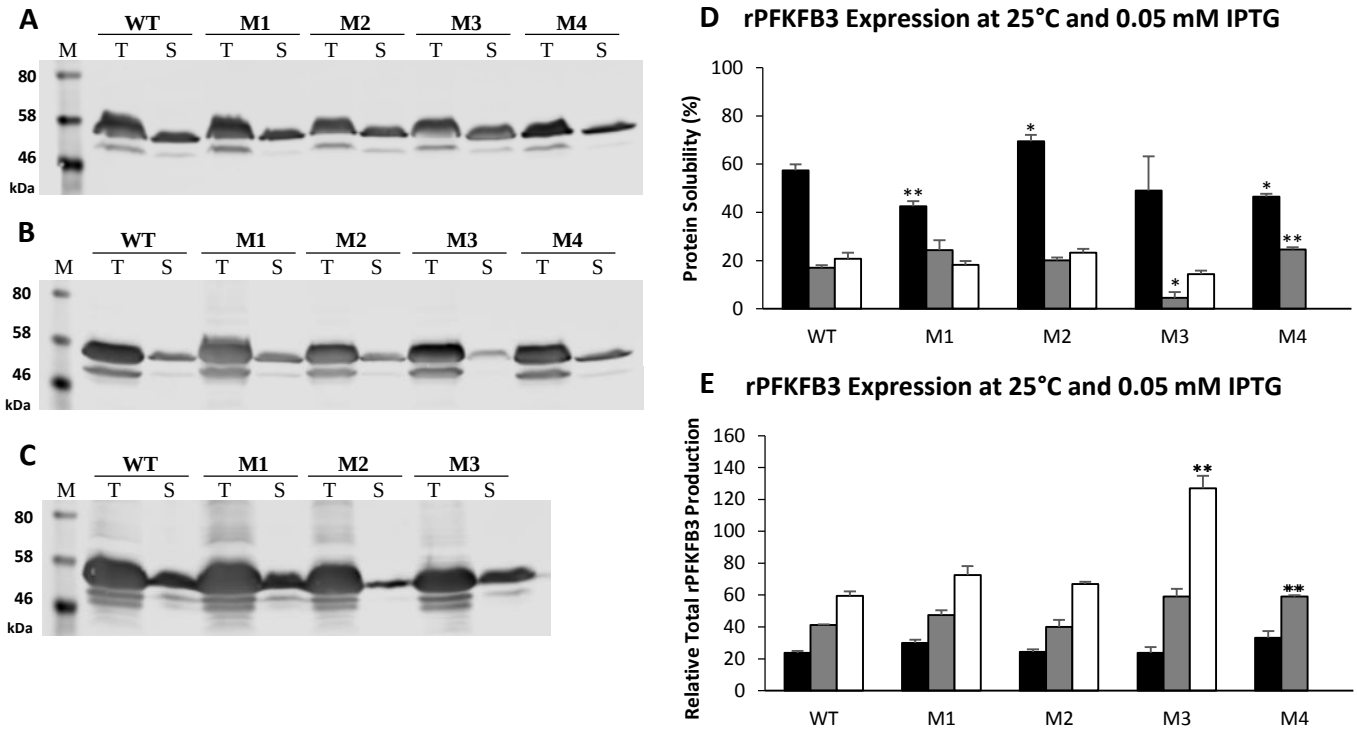


Fig. 3.3. (**A-C**) Western blot of rPFKFB3 wild-type and variants expression in (**A**) BL21 (DE3) pLysS, (**B**) SHuffle and (**C**) BL21-CodonPlus *E. coli* strains. (**D**) Relative protein solubility percentage between soluble and total fraction was plotted. (**E**) Total rHuEPO production was plotted as arbitrary units. For every *E. coli* strain (BL21 (DE3) pLysS [■], SHuffle [▩] and BL21-CodonPlus [□]), triplicate biological replicates were performed for data generation and error bars represent the ± SE of the mean; statistically significant difference was performed using a two-sided unpaired t-test (*P < 0.05, ** P <0.01); M, prestained SDS-PAGE marker (BIO-RAD); T, Total fraction; S, Soluble fraction;

TABLE 3.1. PFKFB3 variants and solubilities profile.

PFKFB3	Aim	Solubility normalised to WT		
		BL21 (DE3) pLysS	BL21-CodonPlus	SHuffle
Wild-type		1.00 ± 0.04	1.00 ± 0.06	1.00 ± 0.13
M1 (W13Y, V14N, V16K)	Increase polarity (see Fig. 3.1)	0.74 ± 0.04	0.88 ± 0.07	1.43 ± 0.24
M2 (Q100E, L103E, A104E)	Diminished positively-charged patch	1.21 ± 0.05	1.12 ± 0.08	1.18 ± 0.07
M3 (N91D, E92Q, N178D)	Diminished local flexibility	0.85 ± 0.25	0.69 ± 0.07	0.26 ± 0.14
M4 (R427D)	Diminished positively-charged patch	0.81 ± 0.02		1.44 ± 0.05

Triplicate biological replicates were performed for each *E. coli* strain for data generation and deviation represent the ± SEM.

3.3.2 *Enhanced surface charge by diminishing the largest positively-charged patch*

A correlation between positively-charged patches and insolubility, that our group observed recently (Chan et al., 2013), is novel. The current work seeks to test this correlation with targeted mutagenesis. Looking graphically at the surface (Fig. 3.4A), there are roughly two large positive (blue) regions, one which stretches between and across the two ligand binding sites, and one which lies around from the ATP binding site. Basic residues around here that are not bind directly to ATP include K96, K99, R107 and K110 along the helix and R132 on a neighbouring helix. Looking at the K96 and the helix in a multiple sequence alignment, many of these charges residues are not conserved (Fig. 3.2E), providing an opportunity for decreasing the positive charge of this patch. The mutations Q100E, L103E, A104E (PFKFB3 M2) were made after structural analysis and posQ calculation in Chan and Warwicker software. The numerical output of this prediction is relatively high for all PFKFB3 proteins, considering the predictor's threshold (above 1.0 predicting a tendency towards insoluble) (Table 3.2). The continuous patch on the opposite side is localised the F6P binding region (2-Pase). In this area the best candidate amino acid to be mutated is R427 (PFKFB3 M4). The overall predicted expression in *E. coli* was insoluble for all the constructs.

PFKFB3 wild-type, M2 and M4 were expressed in the three *E.coli* mentioned above. As shown in Fig. 3.3D mutants M2 and M4 enhanced protein solubility compared to wild-type in BL21 (DE3) pLysS and SHuffle, respectively. On the other hand, significant decreased in M4 solubility in BL21 (DE3) pLysS was obtained. No significant change was observed for M2 in SHuffle and BL21-CodonPlus. In addition, Fig. 3.3E shows no significant change was achieved for M2 and M4 in protein expression among the bacteria expression systems, except for M4 in SHuffle.

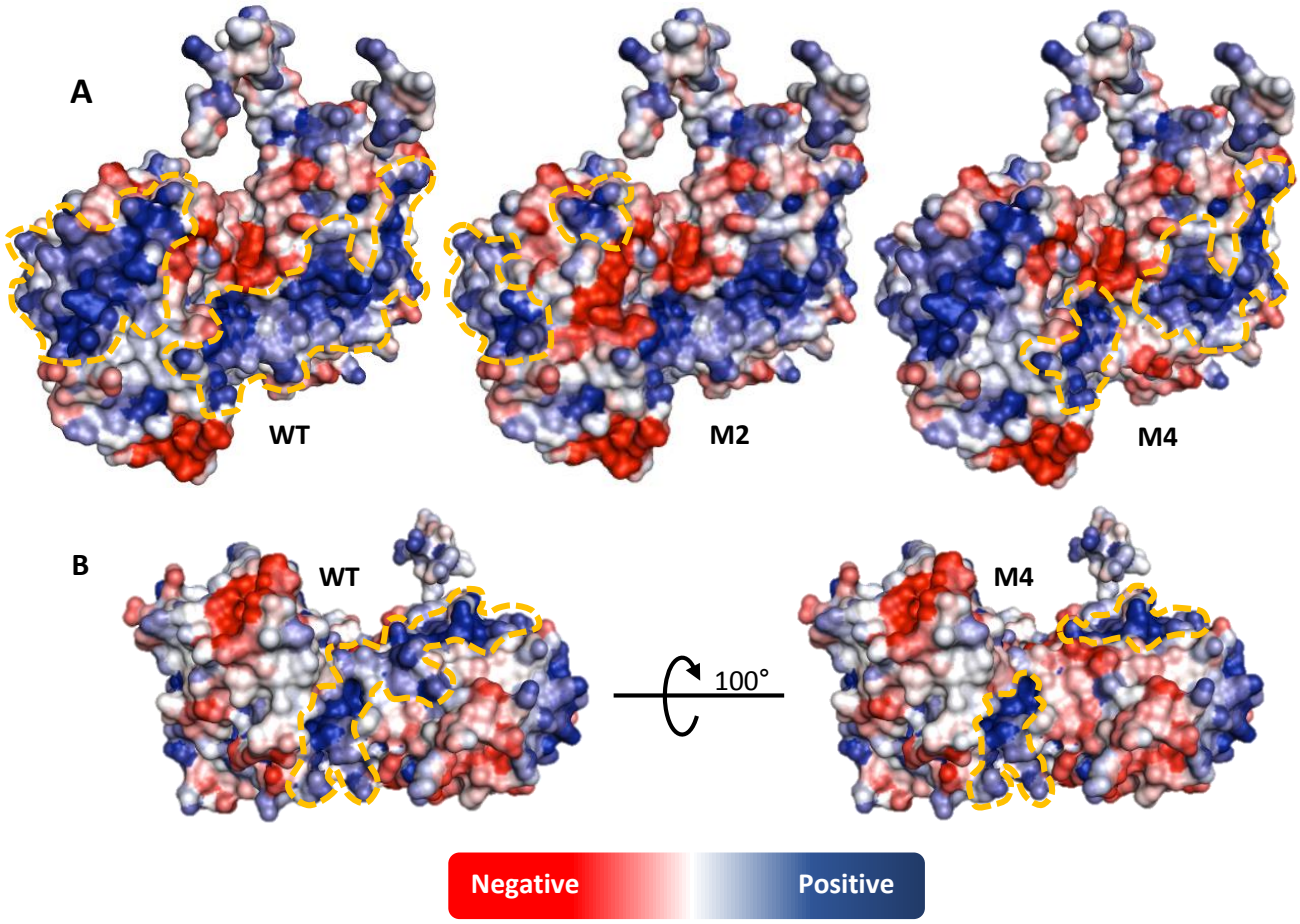


Fig. 3.4. PFKFB3 wild-type, M2 and M4 showing the electrostatic potential patches on the surface (PDB ID: 2AXN). (**A**) Lateral view showing the two largest positively-charged patches. (**B**) Bottom view showing the breakage of the largest positively-charged patch in mutant M4.

TABLE 3.2. Computational calculations on PFKFB3 structure

Protein	Aim	posQ ratio to threshold
rPFKFB3 wild-type		4.46
rPFKFB3 M1 (W13Y, V14N, V16K)	Increase polarity (see Fig. 3.1)	4.47
rPFKFB3 M2 (Q100E, L103E, A104E)	Diminished positively-charged patch	4.02
rPFKFB3 M3 (N91D, E92Q, N178D)	Diminished local flexibility	4.48
rPFKFB3 M4 (R427D)	Diminished positively-charged patch	2.70

Those proteins with posQ ratio above 1.0 are predicted as insoluble and below 1.0 as soluble.

3.3.3 *Improving stability by adding charged helical capping residues: Enhancing PFKFB3 stability*

Poor local stability in a protein can lead to peptide flexibility, partial unfolding, and aggregation (Hwang and Park, 2008). Increasing the structural stability of flexible regions that could partially unfold and thereby lead to exposed non-polar regions and reduced solubility is another consensus view of how to increase solubility (Malakauskas and Mayo, 1998, Dantas et al., 2003, Fowler et al., 2005). Fig. 3.5 shows the B-factors in crystal structures in order to spot the warmer (red) colours that are more flexible. On the other hand, avoiding the cooler (blue) colours that are more ordered (presumed more stable). The more flexible regions could be the source of unfolding fluctuations. Once partially unfolded, exposing non-polar regions, aggregation may be more likely. In the inducible structure, a small area is missing (E446 to P452) followed by most of the C-terminal section (from V461). Also, this area is non-ordered and partially unfolded. In order to have a better understanding of the thermostability distribution in the structure, the liver PFKFB homologue was used to compare these locations (Fig. 3.5A-B). We then applied substitutions to Q92 and D178 as charged helical caps on the highlighted regions around the loop area R88-E92 and K175-N178 (Fig. 3.5). These two areas have shown to be the most flexible on both structures.

We have observed no enhancement in PFKFB3 M3 solubility when expressed under the expression conditions mentioned in section 3.2.3. Furthermore, when expressed in SHuffle, PFKFB3 M3 had significantly lower solubility than the wild-type (Fig. 3.3D). On the other hand, when expressed in BL21-CodonPlus 2-fold of the protein expression is achieved. In contrast, no significant modification in the other two *E. coli* strains compared to the wild-type (Fig. 3.3E).

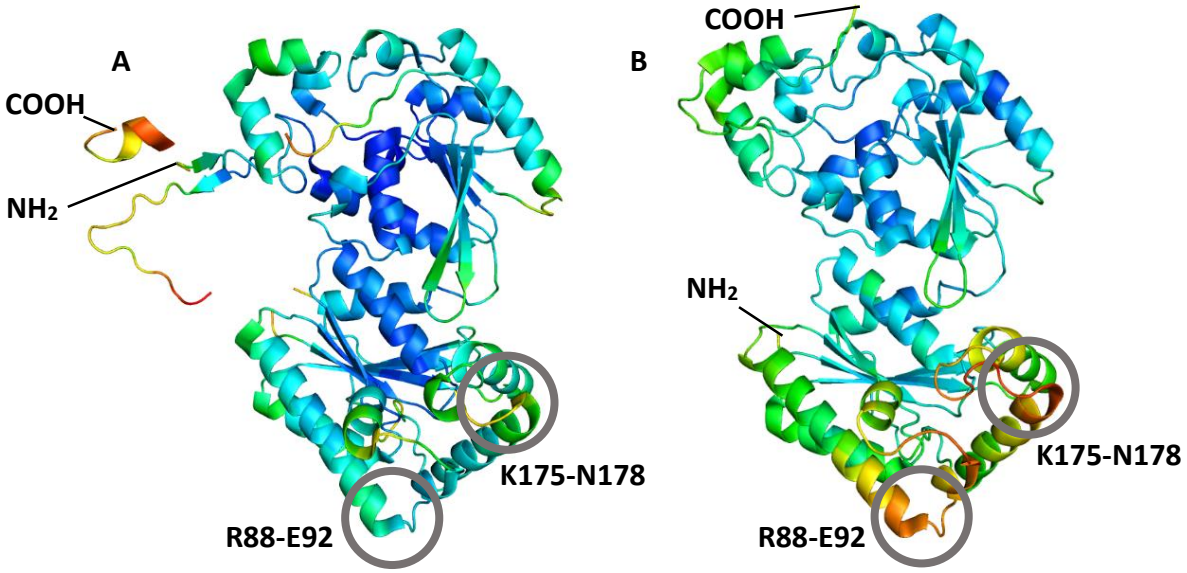


Fig. 3.5. Tertiary structure cartoon representation of thermal stability by B–factor spectrum. The cooler colours are more ordered and warmer colours are more flexible. (**A**) Crystal structure of the inducible form PFKFB3 (2AXN) and (**B**) liver form PFKFB1 (1K6M).

3.4 Discussion

Protein features that alter solubility and aggregation are likely to be environment and protein dependent. In the current study we look at three different predicted features, positive charge patches, non-polar patches, and polypeptide flexibility. It is unlikely, *a prior*, that all three features will have similar effects on protein solubility in a given environment. A key feature in this study is the diminish of positive charge patches since PFKFB3 exhibits a relatively large maximal positive patch in comparison with a threshold derived in a recent study (Chan et al., 2013). Furthermore, larger proteins (such as PFKFB3, 520 amino acids) can lead to more opportunity for aggregation-prone features to appear (Niwa et al., 2009). Additionally, this protein is difficult to express in *E. coli* (McKenzie E.A., unpublished data), and it is to expression systems (rich in nucleic acid) that the positive patch threshold applies. Amongst substantial variation in the solubility results, between mutants and expression systems, one consistent result is the increase in solubility in all expression systems for mutant M2, one of the two aimed at reducing positive patch size. The mutant types are now discussed, in turn.

We calculated the non-polar patch (hydrophobicity) on the monomer structure of PFKFB3. The distribution of the most non-polar patch is localised along the poorly conserved amino-terminal region. On this area, PFKFB3 M1 comprises a less non-polar patch compared to the wild-type by mutating W13Y, V14N and V16K. On the one hand W13 is relatively unusual in exposing a tryptophan side-chain on the surface, such as this case. Through comparison to the dimer of a homologue (1K6M), it appears that the N-terminal region could be proximal to monomer-monomer interactions, but the lack of conservation at the mutated sites of M1 argue against those specific wild-type amino acids being functionally crucial. The results, show variability in solubility relative to PFKFB3 wild-type for M1 in the expression systems (Fig. 3.3D). The lack of a consistent increase in solubility suggests that the non-polar

region of M1 is not generally key to aggregate formation. Other non-polar regions may be important. It is interesting that solubility does increase in the SHuffle system. Perhaps where there is a larger fraction of natively folded protein, shielding interior non-polar regions (as may be the case for SHuffle), then the non-polar exterior plays more of a role in mediating protein-protein interactions.

Secondly, we applied the algorithm to compute the charged surface of PFKFB3 (Fig. 3.4). The biggest positive patch, 4.46 centred between amino acids 122 and 398.This patch seems to wrap around the protein surface (Fig. 3.4 A-B WT). Also, this patch covers the region on ADP binding domain (2-Kase), but facing exterior, not binding directly (Supplementary Fig. S3.1). The substitutions around here diminish the overall patch (PFKFB3 M1: Q110E, L103E and A104E). On the opposite side is localised the F6P binding region (2-Pase), where the target residue to be mutated is the poorly conserved R427 (PFKFB3 M4). This gives a significant decrease in largest positive patch, compared to the wild-type (Table 3.2). Whereas the solubility of M4 relative to wild-type depends on expression system, that for M2 is consistently elevated in all three systems. The increases in solubility for M2 are relatively small, but perhaps indicative that positive charge is playing a role in these expression systems. If this effect were in line with the published hypothesis (Chan et al., 2013), then positive charge on the proteins would be interacting with negative charges (possibly mRNAs) as transient intermediates on the way to aggregate formation. Other interpretations could be made for a role of positive charge in aggregation, with literature observations that carboxyl groups of aspartic and glutamic acid bind to water more strongly than do amino and guanidine groups of arginine and lysine (Kuntz, 1971, Collins and Washabaugh, 1985, Collins, 1997, Kramer et al., 2012, Chong and Ham, 2014).

Lastly, in PyMOL, we generated symmetry mates in the crystal to look at packing and possible influence on flexibility. In addition, we have analysed dimerisation based on the

available homologue dimer structure (1K6M) as a template, aligning to generate a PFKFB3 dimer with the same coordinate origin as 2AXN. This analysis allowed us to suggest regions for stabilisation, based on high crystallographic B-factors (Fig. 3.5), in particular introducing helical capping interactions. However, the predicted decrease in local flexibility did not improved solubility among *E. coli* strains, rather it consistently reduced solubility. Possibly flexibility in this region promotes folding and acquiring native structure i.e. perhaps the folding process in addition to the native state stability should be considered.

In conclusion, as might be anticipated, the different predicted modes of potential effect on solubility lead to different results. Variations between wild-type and mutant solubilities are mostly rather small, preventing clear-cut conclusions. It does appear, however, that charge properties (positive charge patches) should be considered alongside non-polar surface and folded state stability, at least in expression systems. The situation may well be different in purified solutions of proteins, without negatively-charged macroions, but in for expression of PFKFB3 mutants in three different *E. coli* expression systems, only a positive patch reduction mutant gave consistently increased solubility.

3.5 Supplementary data

Residue			Ratio				
LEU	2	HYD		ARG	107	POS	4.465
GLU	3	HYD		ASP	108	NEG	
LEU	4	HYD		VAL	109	NEG	
THR	5	NEG		LYS	110	POS	4.465
GLN	6	HYD		SER	111	POS	4.465
SER	7	HYD		TYR	112	NEG	
ARG	8	NEG		LEU	113	NEG	
VAL	9	POS	0.092	ALA	114	NEG	
GLN	10	POS	0.008	LYS	115	HYD	
LYS	11	NEG		GLU	116	NEG	
ILE	12	HYD		GLY	117	NEG	
TRP	13	HYD		GLY	118	NEG	
VAL	14	HYD		GLN	119	NEG	
PRO	15	HYD		ILE	120	NEG	
VAL	16	POS	0.013	ALA	121	NEG	
ASP	17	HYD		VAL	122	POS	4.465
HIS	18	HYD		PHE	123	POS	4.465
ARG	19	HYD		ASP	124	POS	4.465
PRO	20	HYD		ALA	125	POS	4.465
SER	21	HYD		THR	126	POS	4.465
LEU	22	HYD		ASN	127	POS	4.465
PRO	23	HYD		THR	128	POS	4.465
ARG	24	HYD		THR	129	POS	4.465
SER	25	NEG		ARG	130	HYD	
CYS	26	HYD		GLU	131	POS	4.465
GLY	27	HYD		ARG	132	POS	4.465
PRO	28	HYD		ARG	133	HYD	
ASN	32	POS	4.465	HIS	134	HYD	
SER	33	HYD		MET	135	POS	4.465
PRO	34	NEG		ILE	136	POS	4.465
THR	35	POS	4.465	LEU	137	POS	4.465
VAL	36	POS	4.465	HIS	138	POS	4.465
ILE	37	POS	4.465	PHE	139	POS	4.465
VAL	38	POS	4.465	ALA	140	POS	4.465
MET	39	POS	4.465	LYS	141	HYD	
VAL	40	POS	0.015	GLU	142	HYD	
GLY	41	POS	0.015	ASN	143	HYD	
LEU	42	POS	0.015	ASP	144	HYD	
PRO	43	POS	4.465	PHE	145	HYD	
ALA	44	POS	4.465	LYS	146	HYD	
ARG	45	POS	4.465	ALA	147	POS	4.465
GLY	46	POS	4.465	PHE	148	POS	4.465
LYS	47	POS	4.465	PHE	149	POS	4.465
THR	48	POS	4.465	ILE	150	POS	4.465
TYR	49	POS	4.465	GLU	151	POS	0.015
ILE	50	POS	4.465	SER	152	POS	4.465
SER	51	HYD		VAL	153	POS	0.015
LYS	52	HYD		CYS	154	POS	0.015
LYS	53	POS	4.465	ASP	155	HYD	
LEU	54	POS	4.465	ASP	156	NEG	
THR	55	HYD		PRO	157	HYD	
ARG	56	NEG		THR	158	HYD	
TYR	57	NEG		VAL	159	POS	4.465
LEU	58	HYD		VAL	160	POS	0.015
ASN	59	HYD		ALA	161	HYD	
TRP	60	NEG		SER	162	HYD	
ILE	61	HYD		ASN	163	POS	4.465
GLY	62	HYD		ILE	164	POS	4.465
VAL	63	HYD		MET	165	HYD	
PRO	64	HYD		GLU	166	HYD	
THR	65	NEG		VAL	167	POS	4.465
LYS	66	HYD		LYS	168	POS	4.465
VAL	67	HYD		ILE	169	NEG	
PHE	68	HYD		SER	170	HYD	
ASN	69	POS	0.001	SER	171	POS	4.465
VAL	70	POS	4.465	PRO	172	POS	4.465
GLY	71	POS	4.465	ASP	173	POS	4.465
GLU	72	HYD		TYR	174	NEG	
TYR	73	HYD		LYS	175	NEG	
ARG	74	POS	4.465	ASP	176	NEG	
ARG	75	POS	4.465	CYS	177	NEG	
GLU	76	HYD		ASN	178	HYD	
ALA	77	HYD		SER	179	POS	0.006
VAL	78	HYD		ALA	180	HYD	
LYS	79	HYD		GLU	181	NEG	
GLN	80	HYD		ALA	182	NEG	
TYR	81	POS	0.093	MET	183	NEG	
SER	82	NEG		ASP	184	NEG	
SER	83	NEG		ASP	185	NEG	
TYR	84	NEG		PHE	186	NEG	
ASN	85	POS	0.093	MET	187	HYD	
PHE	86	POS	0.093	LYS	188	POS	0.042
PHE	87	POS	4.465	ARG	189	HYD	
ARG	88	HYD		ILE	190	NEG	
PRO	89	POS	4.465	SER	191	NEG	
ASP	90	HYD		CYS	192	HYD	
ASN	91	HYD		TYR	193	POS	4.465
GLU	92	HYD		GLU	194	NEG	
GLU	93	HYD		ALA	195	HYD	
ALA	94	HYD		SER	196	POS	4.465
MET	95	POS	4.465	TYR	197	NEG	
LYS	96	POS	4.465	GLN	198	POS	4.465
VAL	97	HYD		PRO	199	HYD	
ARG	98	POS	4.465	LEU	200	HYD	
LYS	99	POS	4.465	ASP	201	NEG	
GLN	100	POS	4.465	PRO	202	NEG	
CYS	101	POS	4.465	ASP	203	NEG	
ALA	102	POS	4.465	LYS	204	HYD	
LEU	103	POS	4.465	CYS	205	HYD	
ALA	104	HYD		ASP	206	NEG	
ALA	105	HYD		ARG	207	POS	4.465
LEU	106	POS	4.465	ASP	208	HYD	
LEU	209	POS	4.465	ARG	311	HYD	
SER	210	POS	4.465	LEU	312	HYD	
LEU	211	POS	4.465	PRO	313	HYD	
ILE	212	HYD		TYR	314	NEG	
LYS	213	POS	0.015	GLU	315	NEG	
VAL	214	POS	4.465	GLN	316	HYD	
ILE	215	POS	4.465	TRP	317	POS	0.06
ASP	216	NEG		LYS	318	POS	0.06
VAL	217	POS	4.465	ALA	319	POS	0.06
GLY	218	HYD		LEU	320	POS	0.06
ARG	219	HYD		ASN	321	HYD	
ARG	220	HYD		GLU	322	HYD	
PHE	221	HYD		ILE	323	NEG	
LEU	222	HYD		ASP	324	NEG	
VAL	223	HYD		ALA	325	POS	4.465
ASN	224	HYD		GLY	326	NEG	
ARG	225	HYD		VAL	327	NEG	
VAL	226	POS	0.015	CYS	328	NEG	
GLN	227	HYD		GLU	329	NEG	
ASP	228	HYD		GLU	330	NEG	
HIS	229	HYD		LEU	331	NEG	
ILE	230	POS	4.465	THR	332	POS	4.465
GLN	231	HYD		TYR	333	POS	4.465
SER	232	HYD		GLU	334	POS	4.465
ARG	233	HYD		GLU	335	NEG	
ILE	234	HYD		ILE	336	POS	4.465
VAL	235	HYD		ARG	337	HYD	
TYR	236	HYD		ASP	338	NEG	
TYR	237	HYD		THR	339	NEG	
LEU	238	POS	4.465	TYR	340	HYD	
MET	239	HYD		PRO	341	HYD	
ASN	240	POS	4.465	GLU	342	HYD	
ILE	241	POS	4.465	GLU	343	POS	4.465
HIS	242	POS	4.465	TYR	344	POS	4.465
VAL	243	POS	4.465	ALA	345	NEG	
GLN	244	POS	4.465	LEU	346	HYD	
PRO	245	POS	4.465	ARG	347	POS	4.465
ARG	246	HYD		GLU	348	POS	4.465
THR	247	POS	4.465	GLN	349	HYD	
ILE	248	POS	4.465	ASP	350	HYD	
TYR	249	POS	0.017	LYS	351	POS	0.006
LEU	250	POS	0.005	TYR	352	POS	0.006
CYS	251	POS	0.005	TYR	353	POS	0.008
ARG	252	POS	0.005	TYR	354	POS	0.008
HIS	253	POS	4.465	ARG	355	POS	0.008
GLY	254	POS	4.465	TYR	356	POS	0.064
GLU	255	POS	4.465	PRO	357	HYD	
ASN	256	POS	4.465	THR	358	HYD	
GLU	257	NEG		GLY	359	NEG	
HIS	258	POS	4.465	GLU	360	NEG	
ASN	259	POS	4.465	SER	361	POS	0.008
LEU	260	HYD		TYR	362	POS	0.008
GLN	261	HYD		GLN	363	HYD	
GLY	262	POS	4.465	ASP	364	NEG	
ARG	263	POS	4.465	LEU	365	NEG	
ILE	264	POS	4.465	VAL	366	HYD	
GLY	265	HYD		GLN	367	NEG	
GLY	266	NEG		ARG	368	HYD	
ASP	267	POS	4.465	LEU	369	NEG	
SER	268	HYD		GLU	370	HYD	
GLY	269	POS	4.465	PRO	371	HYD	
LEU	270	HYD		VAL	372	NEG	
SER	271	POS	4.465	ILE	373	HYD	
SER	272	POS	4.465	MET	374	NEG	
ARG	273	POS	4.465	GLU	375	NEG	
GLY	274	POS	4.465	LEU	376	POS	4.465
LYS	275	POS	4.465	GLU	377	POS	4.465
LYS	276	POS	4.465	ARG	378	POS	4.465
PHE	277	HYD		GLN	379	HYD	
ALA	278	POS	4.465	GLU	380	POS	0.052
SER	279	POS	4.465	ASN	381	POS	0.052
ALA	280	POS	4.465	VAL	382	POS	0.052
LEU	281	HYD		LEU	383	POS	0.017
SER	282	NEG		VAL	384	POS	0.017
LYS	283	HYD		ILE	385	POS	0.017
PHE	284	HYD		CYS	386	POS	4.465
VAL	285	POS	0.017	HIS	387	POS	4.465
GLU	286	NEG		GLN	388	POS	4.465
GLU	287	HYD		ALA	389	POS	4.465
GLN	288	HYD		VAL	390	POS	4.465
ASN	289	HYD		LEU	391	POS	0.005
LEU	290	HYD		ARG	392	POS	4.465
LYS	291	HYD		CYS	393	NEG	
ASP	292	NEG		LEU	394	NEG	
LEU	293	NEG		LEU	395	POS	4.465
ARG	294	HYD		ALA	396	NEG	
VAL	295	HYD		TYR	397	NEG	
TRP	296	NEG		PHE	398	POS	4.465
THR	297	HYD		LEU	399	POS	4.465
SER	298	HYD		ASP	400	NEG	
GLN	299	HYD		LYS	401	NEG	
LEU	300	NEG		SER	402	NEG	
LYS	301	HYD		ALA	403	NEG	
SER	302	POS	4.465	GLU	404	NEG	
THR	303	POS	4.465	GLU	405	POS	4.465
ILE	304	HYD		MET	406	POS	4.465
GLN	305	HYD		PRO	407	POS	0.006
THR	306	HYD		TYR	408	POS	4.465
ALA	307	NEG		LEU	409	POS	4.465
GLU	308	NEG		LYS	410	POS	4.465
ALA	309	HYD		CYS	411	POS	0.005
LEU	310	NEG		PRO	412	POS	4.465
LEU	413	POS	4.465				
HIS	414	POS	4.465				
THR	415	POS	0.005				
VAL	416	POS	0.005				
LEU	417	POS	0.005				
LYS	418	HYD					
LEU	419	POS	4.465				
THR	420	POS	4.465				
PRO	421	POS	4.465				
VAL	422	HYD					
ALA	423	HYD					
TYR	424	POS	4.465				
GLY	425	POS	4.465				
CYS	426	POS	4.465				
ARG	427	POS	4.465				
VAL	428	POS	4.465				
GLU	429	POS	4.465				
SER	430	HYD					
ILE	431	HYD					
TYR	432	HYD					
LEU	433	HYD					
ASN	434	HYD					
VAL	435	HYD					
GLU	436	POS	4.465				
SER	437	HYD					
VAL	438	POS	4.465				
CYS	439	POS	4.465				
THR	440	POS	4.465				
HIS	441	POS	4.465				
ARG	442	POS	4.465				
GLU	443	POS	4.465				
ARG	444	HYD					
SER	445	HYD					
ASN	453	HYD					
PRO	454	POS	0.224				
LEU	455	POS	0.025				
MET	456	HYD					
ARG	457	HYD					
ARG	458	HYD					
ASN	459	HYD					
SER	460	HYD					
SER	731	HYD					
SER	732	HYD					

Table S3.1. Solubility profile of rPFKFB3 WT from the charged patch calculator. Complete posQ ratio output for rPFKFB3 structure (PDB ID: 2AXN) is shown. The largest positive patches are represented by blue (ratio > 1.0). Those proteins with ratio above 1.0 are predicted as insoluble and below 1.0 as soluble. Ratio: largest positively-charged patch (posQ) value from the charged patch calculator (Chan et al., 2013). The charged mutations proposed in this study are highlighted in red. Charge patches: HYD, hydrophobic (non-charged); NEG, negatively-charged; POS, positively-charged.

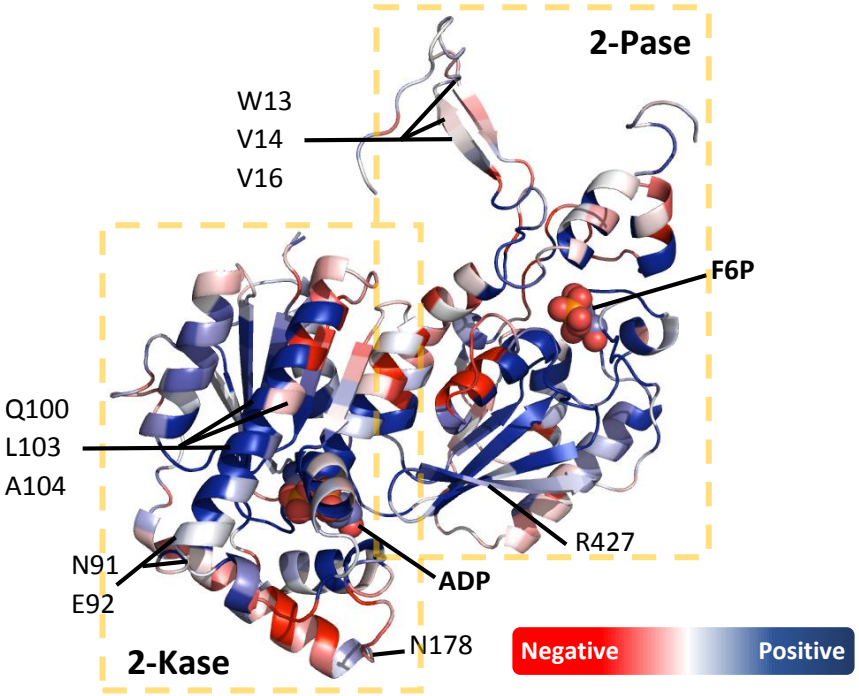


Figure S3.1. PFKFB3 WT tertiary structure ribbon representation of the electrostatic potential distribution and the two binding domains. Overall localisation of all the target residues of the PFKFB3 variants. Charge distribution calculated using the algorithm developed in our group (Chan et al., 2013).

Acknowledgements

We would like to thank Dr. E. Ceh and Dr. J. Suarez for supply of the *Escherichia coli* strains, and CONACyT for contributing PhD funds.

3.6 References

ASHKENAZY, H., EREZ, E., MARTZ, E., PUPKO, T. & BEN-TAL, N. 2010. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res,* 38**,** W529-33.

BAZAN, J. F., FLETTERICK, R. J. & PILKIS, S. J. 1989. Evolution of a bifunctional enzyme: 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase. *Proc Natl Acad Sci U S A,* 86**,** 9642-6.

BONDOS, S. E. & BICKNELL, A. 2003. Detection and prevention of protein aggregation before, during, and after purification. *Anal Biochem,* 316**,** 223-31.

BROOKE, D. G., VAN DAM, E. M., WATTS, C. K., KHOURY, A., DZIADEK, M. A., BROOKS, H., GRAHAM, L. J., FLANAGAN, J. U. & DENNY, W. A. 2014. Targeting the Warburg Effect in cancer; relationships for 2-arylpyridazinones as inhibitors of the key glycolytic enzyme 6-phosphofructo-2-kinase/2,6-bisphosphatase 3 (PFKFB3). *Bioorg Med Chem,* 22**,** 1029-39.

CAVALIER, M. C., KIM, S. G., NEAU, D. & LEE, Y. H. 2012. Molecular basis of the fructose-2,6-bisphosphatase reaction of PFKFB3: transition state and the C-terminal function. *Proteins,* 80**,** 1143-53.

CELNIKER, G., NIMROD, G., ASHKENAZY, H., GLASER, F., MARTZ, E., MAYROSE, I., PUPKO, T. & BEN-TAL, N. 2013. ConSurf: Using Evolutionary Data to Raise Testable Hypotheses about Protein Function. *Israel Journal of Chemistry,* 53**,** 199-206.

CHAN, P., CURTIS, R. A. & WARWICKER, J. 2013. Soluble expression of proteins correlates with a lack of positively-charged surface. *Sci Rep,* 3**,** 3333.

CHEETHAM, J. C., SMITH, D. M., AOKI, K. H., STEVENSON, J. L., HOEFFEL, T. J., SYED, R. S., EGRIE, J. & HARVEY, T. S. 1998. NMR structure of human erythropoietin and a comparison with its receptor bound conformation. *Nat Struct Biol,* 5**,** 861-6.

CHESNEY, J., MITCHELL, R., BENIGNI, F., BACHER, M., SPIEGEL, L., AL-ABED, Y., HAN, J. H., METZ, C. & BUCALA, R. 1999. An inducible gene product for 6-phosphofructo-2-kinase with an AU-rich instability element: Role in tumor cell glycolysis and the Warburg effect. *Proceedings of the National Academy of Sciences,* 96**,** 3047-3052.

CHONG, S. H. & HAM, S. 2014. Interaction with the surrounding water plays a key role in determining the aggregation propensity of proteins. *Angew Chem Int Ed Engl,* 53**,** 3961-4.

CLEM, B., TELANG, S., CLEM, A., YALCIN, A., MEIER, J., SIMMONS, A., RASKU, M. A., ARUMUGAM, S., DEAN, W. L., EATON, J., LANE, A., TRENT, J. O. & CHESNEY, J. 2008. Small-molecule inhibition of 6-phosphofructo-2-kinase activity suppresses glycolytic flux and tumor growth. *Mol Cancer Ther,* 7**,** 110-20.

CLEM, B. F., O'NEAL, J., TAPOLSKY, G., CLEM, A. L., IMBERT-FERNANDEZ, Y., KERR, D. A., 2ND, KLARER, A. C., REDMAN, R., MILLER, D. M., TRENT, J. O., TELANG, S. & CHESNEY, J. 2013. Targeting 6-phosphofructo-2-kinase (PFKFB3) as a therapeutic strategy against cancer. *Mol Cancer Ther,* 12**,** 1461-70.

COLLINS, K. D. 1997. Charge density-dependent strength of hydration and biological structure. *Biophys J,* 72**,** 65-76.

COLLINS, K. D. & WASHABAUGH, M. W. 1985. The Hofmeister effect and the behaviour of water at interfaces. *Q Rev Biophys,* 18**,** 323-422.

DANTAS, G., KUHLMAN, B., CALLENDER, D., WONG, M. & BAKER, D. 2003. A Large Scale Test of Computational Protein Design: Folding and Stability of Nine Completely Redesigned Globular Proteins. *Journal of Molecular Biology,* 332**,** 449-460.

DYDA, F., HICKMAN, A., JENKINS, T., ENGELMAN, A., CRAIGIE, R. & DAVIES, D. 1994. Crystal structure of the catalytic domain of HIV-1 integrase: similarity to other polynucleotidyl transferases. *Science,* 266**,** 1981-1986.

EL-MAGHRABI, M. R., NOTO, F., WU, N. & MANES, N. 2001. 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase: suiting structure to need, in a family of tissue-specific enzymes. *Curr Opin Clin Nutr Metab Care,* 4**,** 411-8.

FINK, A. L. 1998. Protein aggregation: folding aggregates, inclusion bodies and amyloid. *Fold Des,* 3**,** R9-23.

FOWLER, S. B., POON, S., MUFF, R., CHITI, F., DOBSON, C. M. & ZURDO, J. 2005. Rational design of aggregation-resistant bioactive peptides: reengineering human calcitonin. *Proc Natl Acad Sci U S A,* 102**,** 10105-10.

GREAVES, R. B. & WARWICKER, J. 2007. Mechanisms for stabilisation and the maintenance of solubility in proteins from thermophiles. *BMC Struct Biol,* 7**,** 18.

GUEX, N. & PEITSCH, M. C. 1997. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis,* 18**,** 2714-23.

HASEMANN, C. A., ISTVAN, E. S., UYEDA, K. & DEISENHOFER, J. 1996. The crystal structure of the bifunctional enzyme 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase reveals distinct domain homologies. *Structure,* 4**,** 1017-29.

HUE, L. & ROUSSEAU, G. G. 1993. Fructose 2,6-bisphosphate and the control of glycolysis by growth factors, tumor promoters and oncogenes. *Advances in Enzyme Regulation,* 33**,** 97-110.

HWANG, I. & PARK, S. 2008. Computational design of protein therapeutics. *Drug Discovery Today: Technologies,* 5**,** e43-e48.

KHAN, M. A., ISLAM, M. M. & KURODA, Y. 2013. Analysis of protein aggregation kinetics using short amino acid peptide tags. *Biochim Biophys Acta,* 1834**,** 2107-15.

KIM, S. G., MANES, N. P., EL-MAGHRABI, M. R. & LEE, Y. H. 2006. Crystal structure of the hypoxia-inducible form of 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase (PFKFB3): a possible new target for cancer therapy. *J Biol Chem,* 281**,** 2939-44.

KRAMER, R. M., SHENDE, V. R., MOTL, N., PACE, C. N. & SCHOLTZ, J. M. 2012. Toward a molecular understanding of protein solubility: increased negative surface charge correlates with increased solubility. *Biophys J,* 102**,** 1907-15.

KUNTZ, I. D. 1971. Hydration of macromolecules. III. Hydration of polypeptides. *Journal of the American Chemical Society,* 93**,** 514-516.

LIN, K., KURLAND, I., XU, L. Z., LANGE, A. J., PILKIS, J., EL-MAGHRABI, M. R. & PILKIS, S. J. 1990. Expression of mammalian liver glycolytic/gluconeogenic enzymes in Escherichia coli: recovery of active enzyme is strain and temperature dependent. *Protein Expr Purif,* 1**,** 169-76.

MAJHI, P. R., GANTA, R. R., VANAM, R. P., SEYREK, E., GIGER, K. & DUBIN, P. L. 2006. Electrostatically driven protein aggregation: beta-lactoglobulin at low ionic strength. *Langmuir,* 22**,** 9150-9.

MALAKAUSKAS, S. M. & MAYO, S. L. 1998. Design, structure and stability of a hyperthermophilic protein variant. *Nat Struct Mol Biol,* 5**,** 470-475.

NASREEN, A., VOGT, M., KIM, H. J., EICHINGER, A. & SKERRA, A. 2006. Solubility engineering and crystallization of human apolipoprotein D. *Protein Sci,* 15**,** 190-9.

NIWA, T., YING, B. W., SAITO, K., JIN, W., TAKADA, S., UEDA, T. & TAGUCHI, H. 2009. Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of Escherichia coli proteins. *Proc Natl Acad Sci U S A,* 106**,** 4201-6.

PATEL, S. B., CAMERON, P. M., FRANTZ-WATTLEY, B., O'NEILL, E., BECKER, J. W. & SCAPIN, G. 2004. Lattice stabilization and enhanced diffraction in human p38 alpha crystals by protein engineering. *Biochim Biophys Acta,* 1696**,** 67-73.

PETI, W. & PAGE, R. 2007. Strategies to maximize heterologous protein expression in Escherichia coli with minimal cost. *Protein Expr Purif,* 51**,** 1-10.

PILKIS, S. J., REGEN, D. M., STEWART, H. B., PILKIS, J., PATE, T. M. & EL-MAGHRABI, M. R. 1984. Evidence for two catalytic sites on 6-phosphofructo-2-kinase/fructose 2,6-bisphosphatase. Dynamics of substrate exchange and phosphoryl enzyme formation. *J Biol Chem,* 259**,** 949-58.

SAKAKIBARA, R., KATO, M., OKAMURA, N., NAKAGAWA, T., KOMADA, Y., TOMINAGA, N., SHIMOJO, M. & FUKASAWA, M. 1997. Characterization of a human placental fructose-6-phosphate, 2-kinase/fructose-2,6-bisphosphatase. *J Biochem,* 122**,** 122-8.

SCHRÖDINGER, L. L. C. 2010. The PyMOL Molecular Graphics System, Version 1.3r1.

WEISS, W. F. T., YOUNG, T. M. & ROBERTS, C. J. 2009. Principles, approaches, and challenges for predicting protein aggregation rates and shelf life. *J Pharm Sci,* 98**,** 1246-77.

YAMAMOTO, T., TAKANO, N., ISHIWATA, K., OHMURA, M., NAGAHATA, Y., MATSUURA, T., KAMATA, A., SAKAMOTO, K., NAKANISHI, T., KUBO, A., HISHIKI, T. & SUEMATSU, M. 2014. Reduced methylation of PFKFB3 in cancer cells shunts glucose towards the pentose phosphate pathway. *Nat Commun,* 5**,** 3480.

Chapter 4

Paper 2:

Increasing solubility in recombinant erythropoietin through modification of surface patches

Carballo-Amador M.A.[12], Warwicker J.[2] and Dickson A.J.[1]

[1] Faculty of Life Sciences, University of Manchester, Michael Smith Building, Oxford Road, Manchester M13 9PT, UK

[2] Faculty of Life Sciences, University of Manchester, Manchester Institute of Biotechnology, 131 Princess Street, Manchester M1 7DN, UK

Abstract

Protein solubility characteristics are important determinants of the success for recombinant proteins in relation to expression, purification, storage and administration. *Escherichia coli* offers a cost-efficient expression system. However, it has an important limitation, whether for biophysical studies or industrial-scale, in the formation of insoluble protein aggregates in the cytoplasm. Several strategies are being implemented to improved soluble expression, ranging from culture conditions to including solubility enhancing tags. Based on protein structure, an algorithm has been developed in our group to predict protein solubility, defining polar and non-polar patches on the protein surface. Using this algorithm, we predicted amino acid changes that could alter expression of forms of recombinant human erythropoietin (rHuEPO) in *E. coli*. Here we report a set of rHuEPO proteins generated by adjusting positively-charged patches with positively- or negatively-charged amino acid mutations: rHuEPO E13K, F48D, R150D, and F48D/R150D. We found that the variant predicted to aggregate (rHuEPO E13K) decreased solubility by 30 to 70% compared to rHuEPO WT. In contrast, those predicted to be soluble (rHuEPO F48D, R150D, and F48D/R150D) increased solubility up to 60% in relation to rHuEPO WT. We therefore found that point mutations verified the predicted effect on rHuEPO solubility.

4.1 Introduction

Biological systems have evolved by orchestration of molecule interactions in order to thrive through evolution. Proteins are key elements in such mechanisms for all known forms of life. In the circulatory system 2.4 million red blood cells are replaced every second in human adults (Sackmann, 1995). This requires highly stable and efficient regulatory mechanisms to fulfil this demand. Erythropoietin, the main glycoprotein behind this task, regulates the growth and proliferation of red blood cell progenitors (Krantz, 1991). Erythropoietin is one of the top-selling therapeutics, with annual sales in excess of $5.77bn in 2013 (Walsh, 2014) and providing a live-saving/benefiting therapy for millions of patients. There remains a continued demand to make erythropoietin at large-scale and to increase the economic efficiency to provide a decrease in cost of goods and, hence, availability to patients. Human erythropoietin (HuEPO) consists of 166 amino acid residues, in a structure that includes three N-linked glycosylation sites (N24, N38 and N83), one O-linked glycosylation (S126) and two disulphide bonds (C7-C161 and C29-C33) (Lai et al., 1986). These complex post-translational modifications (PTMs) are the main challenge for expression of HuEPO in heterologous expression systems, and in the cost-efficient *E. coli* system none of these PTMs are added in the cytoplasmic environment. Using bacteria to produce recombinant proteins efficiently entails a significant challenge since cysteine mispairing may lead to misfolding and low yields (Fink, 1998). In order to overcome this challenge, engineered strains have been developed such as the SHuffle strains from New England Biolabs. In this *E. coli* strain the cytoplasmic environment is altered by the overexpression of DsbC disulphide bond isomerase and by deletion of two reductases (glutaredoxin [*gor*] and thioredoxin [*trxB*]) (Lobstein et al., 2012). In addition, the glycosylation of HuEPO, to contribute ~40% of the overall molecular mass, generates stability and solubility to the molecule as a therapeutic (Davis et al., 1987, Narhi et

al., 1991, Banks, 2011) and the lack of glycosylation of recombinant HuEPO (rHuEPO) derived from *E. coli* leads to aggregation during expression and, potentially, subsequent purification, storage and delivery (Cheetham et al., 1998). This protein aggregation phenomena during expression in *E. coli* leads to incorporation into inclusion bodies (IBs), a consequence of interactions of partially folded, misfolded or unfolded recombinant proteins during protein folding in the cytoplasm (Fink, 1998). The importance of glycans to IBs formation was illustrated by mutation of the three *N*-glycosylation sites on HuEPO to polar residues (N24K, N38K and N83K). This decreased IBs formation and facilitated the purification of quality protein to provide the solution of the rHuEPO crystal structure (Cheetham et al., 1998, Narhi et al., 2001).

Protein aggregation involves reversible and irreversible reactions, such as chemical aggregation (oxidation or covalent bonds, e.g. disulphide bond formation) and/or physical aggregation (non-covalent interactions between hydrophobic surfaces) (Wang, 2005). Several attempts have been made to diminish hydrophobic patches on the surface to prevent protein aggregation (Jenkins et al., 1995, Li et al., 1999, Das and Georgiadis, 2001, Slovic et al., 2003, Fan et al., 2004, Lawson et al., 2009). In this context, Buchanan et al., (2012) generated an improvement in expression, stability and solubility of rHuEPO and granulocyte colony-stimulating factor (G-CSF) by application of the *in vitro* ribosome display technique, in combination with three parallel selection pressures (reducing agent, elevated temperature and hydrophobic interaction chromatography matrices). In the case of rHuEPO, a variant encoding four mutations resulted in a form that was less prone to aggregate (Buchanan et al., 2012). Furthermore, the application of fusion tags to improve rHuEPO solubility has been used successfully (Ahn et al., 2011, Jeong et al., 2014). Some of these fusion partners, including NusA and maltose-binding protein (MBP), have large negatively-charged areas and may be involved in promoting folding of the target protein by limiting protein aggregation (Zhang et

al., 2004). Engineering of negatively-charged areas on protein surfaces is gaining strength as an approach for increasing solubility (Trevino et al., 2007, Perchiacca et al., 2012, Chan et al., 2013, Chong and Ham, 2014).

Here we report a novel experimental approach targeted at improving rHuEPO solubility for expression in *E. coli*. This research arose from a recently published hypothesis from our group, stating that soluble expression of proteins was inversely correlated with the size of the largest positively-charged patch on the protein surface (Chan et al., 2013). Here we test this hypothesis by mapping surface charge of rHuEPO, focusing on adjusting positively-charged patches with positively- or negatively-charged amino acids mutations. Based on this, we have generated a set of mutants ranging from more (rHuEPO E13K) to less positively-charged surface patches (rHuEPO F48D, R150D and F48D/R150D) compared to natural (wild type) rHuEPO (rHuEPO WT). In this study, experimental results correlate with the prediction, i.e. more positively-charged patches on surface leads to protein aggregation in the cytoplasm of *E. coli*. Further application of this approach could offer a powerful tool for the rational design of proteins with enhanced solubility and stability.

4.2 Material and methods

4.2.1 *rHuEPO solubility profile and mutant design*

The PyMOL Molecular Graphics System version 1.3 (Schrödinger, 2010) and Swiss-PdbViewer 4.0.1 (Guex and Peitsch, 1997) were used to analyse rHuEPO structural and sequence features. The protein solubility predictions were calculated using an algorithm developed in our group (Chan et al., 2013) (Supplementary Table S4.1 and S4.2). The

algorithm computes structured-based parameters, including, nonQmax –the maximal non-charged patch-, posQmax –the maximal size of a positively charged patch- and their multiplicative combination, versus thresholds calculated from Niwa dataset of experimental solubilities determined for cell-free expression of *E. coli* proteins (Niwa et al., 2009). These thresholds have been derived by looking for the value of a parameter that best separates less and more soluble proteins, with those proteins with the largest positive patch predicted as least soluble (ratio above 1.0 relative to the threshold). Positive patches with ratio below 1.0 are predicted as more soluble. Current work concentrates on positive patches (posQ), since this structure-based feature gives the best separation in the benchmark calculations (Chan et al., 2013). Based on the rHuEPO wild-type solubility profile, substitutions to modify posQ from the PDB file were carried out in Swiss-PdbViewer by applying the "MUTATE" function. Then, new PDB files were generated for further analysis in the Chan and Warwicker algorithm (Table 4.1). This procedure gave the following candidates to be mutated: rHuEPO E13K, rHuEPO F48D, rHuEPO R150D and the double mutant rHuEPO F48D/R150D. Aspartic acid was selected as the negatively-charged residue to introduce due to its short side chain compared to glutamic acid, lowering the possibility of non-specific interactions with the surrounding side chains. All these calculations were performed using a modified crystal structure of an analogue rHuEPO taken from the Protein Data Bank (PDB ID: 1EER). These modifications consist in the background of the native HuEPO sequence in order to maintain consistency with our experimental and native rHuEPO cDNA (K24N, K38N, K83N, N121P and S122P).

4.2.2 *Surface rHuEPO positively-charged patch distribution and conservation among species*

Multiple sequence alignments and surface mapping coloured by residue conservation were performed using ConSurf with default parameters (Ashkenazy et al., 2010, Celniker et al., 2013). This default parameters consist suing the UniRef90 database (Suzek et al., 2007), which removes redundancy at 90% sequence identity. In addition, the default number of mohomologues sequences was set to 150 with a Position-Specific Iterated BLAST (PSI-BLAST) E-value to 0.0001 to minimise the chance of including non-homologues. In addition, a multiple EPO study was made with 115 models generated from a clustal alignment after searching for EPO orthologues with BLAST and checking that annotation indicates truely EPO in the various species. Structural models were made from the clustal alignment to the known structure, using an automated pipeline developed in our group, and patch calculations made based on the structural models.

4.2.3 *Construction of rHuEPO mutants and expression vectors*

Human erythropoietin cDNA was amplified from a pre-existing mammalian expression vector by applying primers containing the restriction sites 5'-*BamHI* and 3'-*EcoRI*. The PCR fragment (lacking of signal peptide) was subcloned into a pHis vector which was kindly provided by Dr. Edward McKenzie of the University of Manchester. This plasmid is a modified version of the commercial pET-16b vector (Novagen) (Section 9.2). The gene sequence for each plasmid was as follows: 5'-6xHis-Thrombin cleavage site-*BamHI*-rHuEPO-*EcoRI*-3'. rHuEPO mutations were introduced using GENEART Site-Directed Mutagenesis System with the enzyme AccuPrime *Pfx* (Invitrogen).

4.2.4 *Protein expression and solubility assay*

The bacterial cell lines used in this study were *Escherichia coli* BL21 (DE3) pLysS and SHuffle *(New England BioLabs). The bacterial strains were transformed with the pHis-rHuEPO plasmids.* Transformed cells were grown overnight in 5 ml working volume of Luria-Bertani (LB) medium (10g tryptone, 5g yeast extract, 5g NaCl) containing 100 µg/ml ampicillin at 37ºC with shaking at 220 rpm. In addition, BL21 (DE3) pLysS were grown with 50 µg/ml chloramphenicol in order to preserve the pLysS plasmid. Next day, 1 ml of pre-culture was transfer to 50 ml 2% (v/v) LB supplemented with 2% (w/v) glucose with 100 µg/ml ampicillin in 250 ml shake flasks by triplicate biological replicates. The shake flasks were incubated in constant temperature at 25ºC with shaking at 180 rpm. Bacteria were grown to an OD_{600} of approximately 0.6-0.8. Protein expression was induced with 0.05 mM IPTG. After 5 h, cells were centrifuged at 6,500 g for 15 min at 4ºC. Bacterial pellets were suspended in 5 ml of lysis buffer (25 mM Tris pH 7.5, 150 mM NaCl, 1% [v/v] Triton X-100) and were stored at -20ºC until future use. The cells were disrupted by six sonication cycles of 30 s at 20% amplitude and then allowed to cool for 30 s on ice water bath. Separation of soluble and total fractions was performed by centrifugation at 18,000 g for 30 min at 4ºC of 1 ml of each sample from the whole cell lysate. The supernatants were collected and handled as the soluble fraction. Then an additional 1 ml of each lysate sample was processed as the total fraction. rHuEPO solubility was calculated by densitometric ratio of soluble to total fraction.

4.2.5 *SDS-PAGE and Western blot*

Recombinant proteins were separated by sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE) in a 12% (w/v) separating gel with a 5% (w/v) stacking gel using the Mini-PROTEAN Tetra Cell (BioRad). Samples containing equal volumes of protein were subjected to heat at 95°C by 5 min in 6x denaturing buffer (375 mM Tris pH 6.8, 12% [w/v] SDS, 60% [v/v] glycerol, 0.06% [w/v] bromophenol blue, 5.5% [v/v] β-mercaptoethanol). Peptides were separated in electrode running buffer (50 mM Tris, 0.38 M glycine, 0.2% [w/v] SDS) at 60 V until samples migrated into the separating gel and then the voltage was increased to 160 V at room temperature. Proteins were transferred to nitrocellulose membrane surrounded by transfer pads that were soaked into blotting buffer (25 mM Tris, pH 7.4, 0.2 M glycine and 20% [v/v] methanol). The transfer was performed using a transblot semi-dry transfer cell (Bio-Rad) at 15V for 45 min. The membranes were blocked overnight for non-specific binding in blocking buffer (5% [w/v] skimmed milk in TBS-Tween pH 7.4) at 4ºC with shaking. For detection of rHuEPO, a mouse anti-polyHis antibody (Sigma) was diluted 1:5000 in blocking buffer solution and the membrane was incubated in this solution for 2 h at room temperature in agitation. The primary antibody was removed and the membrane was washed three times (5 min each time) at room temperature in TBS-Tween. Then, incubated with an IR-labeled secondary Donkey anti-Mouse IgG antibody (LI-COR) diluted 1:15000 in blocking buffer solution at room temperature for 45 min. Followed the incubation, the secondary antibody was removed and the membrane was washed three times as mention before. For IR detection, blots were imaged with the Odyssey Imaging System. Bands were quantified in Image Studio Lite software (LI-COR).

4.3 Results

4.3.1 *Positively charged patches govern rHuEPO WT surface area*

We applied a recently published algorithm to generate a solubility profile for rHuEPO and identify potential amino acids for mutation to rHuEPO variants with predicted enhanced solubility (Table 4.1). This method (Chan et al., 2013) demonstrated a correlation between positive charge patches and insolubility in a cell-free expression system (Niwa et al., 2009). Design for improved solubility using the algorithm therefore involved identification and reduction of the larger positively-charged patch. From the algorithm prediction we identified the relative insolubility of rHuEPO (rHuEPO wild type, WT). For the protein variants shown in Table 4.1, substitutions R150D and F48D, individually or in combination, showed a lowered positive patch size and were predicted to be more soluble than rHuEPO WT. In contrast, substitution E13K had an increased positive patch than wild type and was predicted to generate a less soluble product. The calculations therefore generated a set of predictions from the relatively insoluble rHuEPO WT to a more insoluble (rHuEPO E13K) or soluble (rHuEPO F48D, R150D and their combination) variants. In addition, the rHuEPO surface charge patches were visualised (Fig. 4.1). Interestingly, the largest positively-charged patch varied substantially through evolution, suggesting (according to the solubility hypothesis being tested) that expression solubility properties of EPO from different species could also be quite divergent (Fig. 4.2). Structure sequence conservation analysis using the ConSurf server showed the highly conserved R150 and the relative lower degree of conservation for the residues E13 and F48 (Fig. 4.3). We next proceeded to analyse the structural implications of these substitutions. The three mutated sites are all on the protein surface, so that a first approximation would

assume minimal disruption to the folded state stability. This assumption is revisited in the Discussion section.

TABLE 4.1. Predicted solubilities of recombinant human erythropoietin.

Protein	Pos patch ratio to threshold	Prediction
rHuEPO wild-type	1.49	Insoluble
rHuEPO F48D	0.75	Soluble
rHuEPO R150D	0.61	Soluble
rHuEPO F48D/R150D	0.47	Soluble
rHuEPO E13K	2.47	Insoluble

Solubility profile was defined as described in Chan et al (2013). Positive (Pos) patch sizes are divided by that best separating soluble and insoluble datasets (Niwa et al., 2009), above 1.0 implies predicted insolubility.

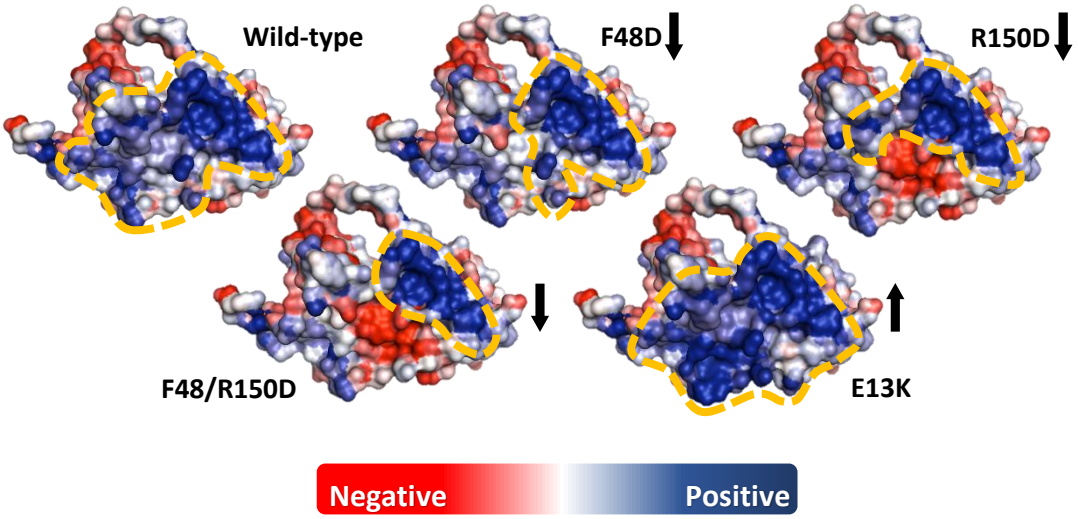


Fig. 4.1. Human erythropoietin wild-type and variants surface illustration showing the electrostatic potential patches (Chan et al., 2013). Amino acids in the largest positive patches are represented by blue, non-charged patches by white and negatively charged by red colour, respectively.

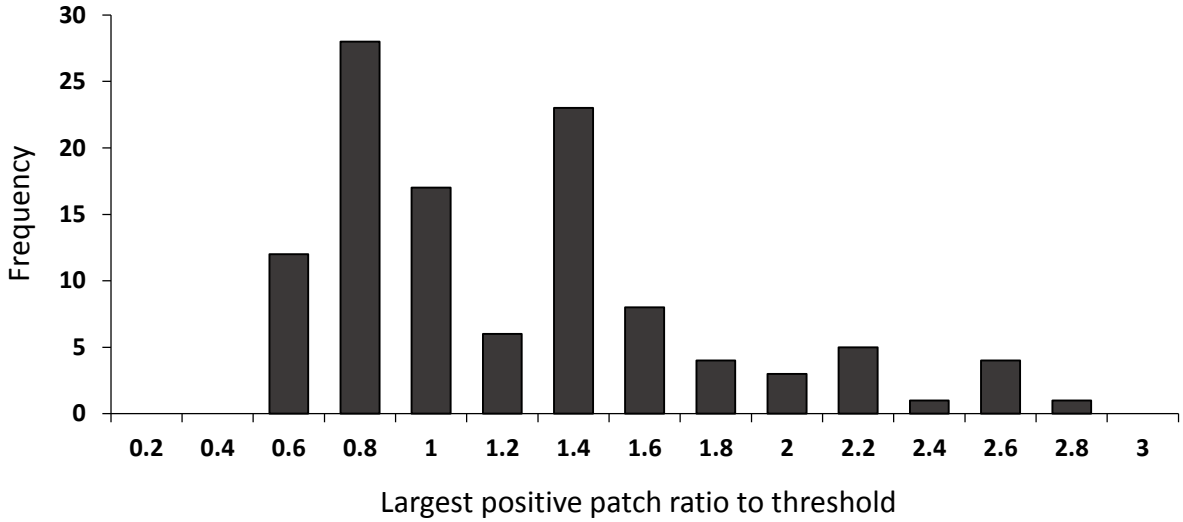


Fig. 4.2. Positively-charged patch on EPO surface through evolution. Positively-charged patches distribution of largest positive patch ratios to the threshold for 115 EPO orthologues, showing that surface charge changes in evolution lead to quite large changes in patch. HuEPO is located roughly in the centre of the distribution according to the calculation of surface charges (positive patch ratio to threshold of 1.49). Frequency is the number of EPO orthologues having a largest positive patch ratio to threshold, in the given x-axis bin.

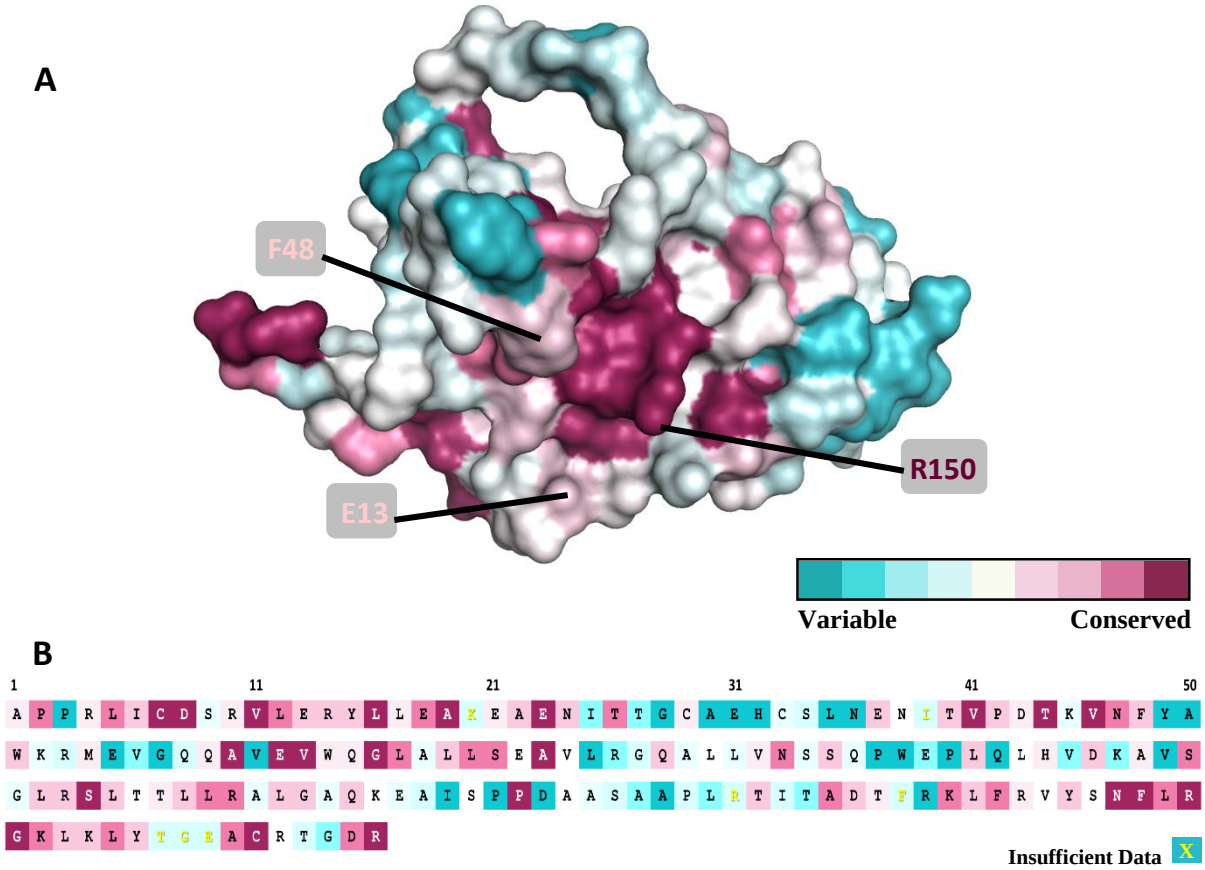


Fig. 4.3. Multiple alignment of HuEPO. (**A**) Sequence and surface map coloured by residue conservation scores of rHuEPO wild-type. The image was rendered using PyMOL. (**B**) Panel shows the conservational result for the input sequence of rHuEPO run with default parameters.

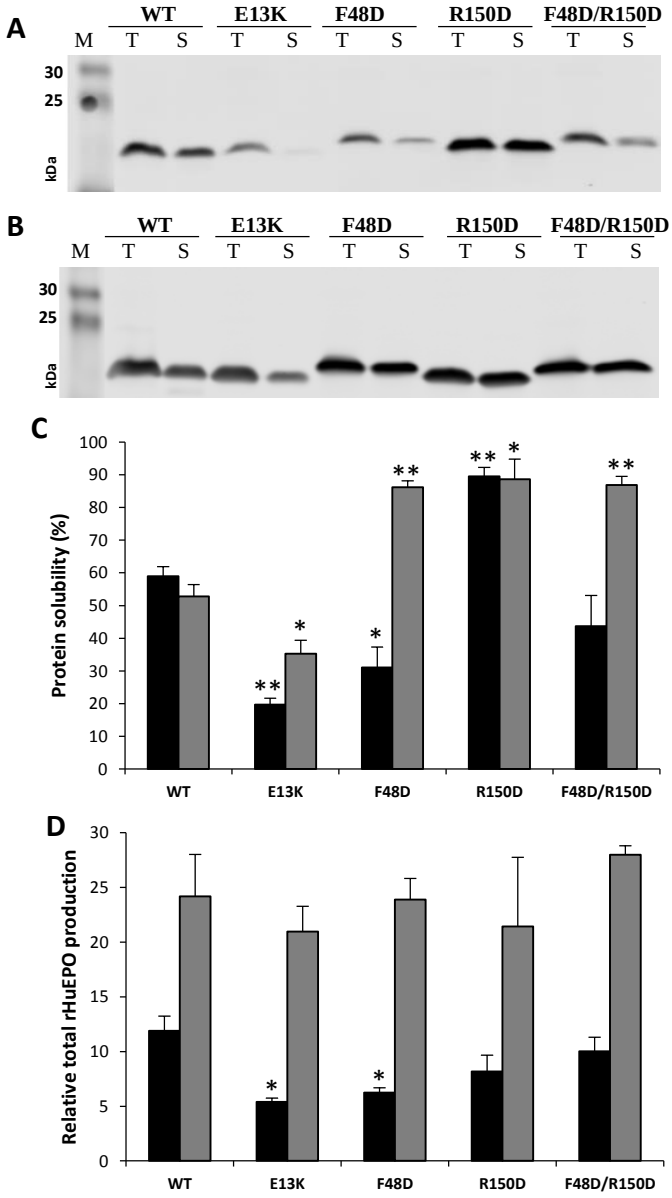


Fig. 4.4. Western blot of rHuEPO expression and solubility degree. (**A-B**) Equal volume of total protein and soluble fraction were probed with Mouse anti-polyHis antibody and were imaged with the Odyssey Imaging System for BL21 (DE3) pLysS (**A**) and SHuffle (**B**) strains. (**C**) Experimental solubility was determined by the distribution of rHuEPO between soluble and inclusion body fraction in *E. coli*. (**D**) Relative total rHuEPO production was plotted as arbitrary units. For every *E. coli* culture, triplicate biological replicates were performed for data generation and error bars represent the ± SEM; statistically significant difference was performed using a two-sided unpaired t-test (*P < 0.05, ** P <0.01). BL21 (DE3) pLysS (■); SHuffle (▣); M, prestained SDS-PAGE marker (Bio-Rad); T, total fraction; S, soluble fraction.

4.3.2 *Enhanced rHuEPO soluble expression*

It has been shown previously that when expressed in *E. coli* rHuEPO WT tended to form insoluble protein aggregates in inclusion bodies (Narhi et al., 2001). The use of sub-optimal temperature and lower inducer concentrations has been argued to be a more appropriate approach to assess protein folding (and hence solubility) of recombinant proteins in *E. coli* (Sevastsyanovich et al., 2009). Use of "low induction" conditions (decreased temperature, lower IPTG challenge) decreased cell growth and the elongation rate of translation (Farewell and Neidhardt, 1998) and induced chaperone activity and protein folding capability (Ferrer et al., 2003). These responses led to better folding and less degradation (Chesshyre and Hipkiss, 1989) and offered a more refined system to investigate the consequences of variant structures in the expressibility and solubility for recombinant protein expression in *E.coli*. Consequently we extended our studies to examine expression and solubility of rHuEPO variants under low induction conditions. Protein solubility agreed with the prediction for BL21 (DE3) pLysS and SHuffle strains with the exceptions in the former strain for rHuEPO F48D and F48D/R150D (Fig. 4.4C). In terms of relative total rHuEPO expression, significant changes among the variants were achieved only for E13Kand F48D compared to the wild-type in the BL21 (DE3) pLysS strain (Fig. 4.4D).

4.4 Discussion

Previously, several research groups have developed methods to predict protein solubility based on sequence (Wilkinson and Harrison, 1991, Davis et al., 1999, Smialowski et al., 2007, Magnan et al., 2009, Price et al., 2011, Huang et al., 2012, Niu et al., 2012, Samak et al., 2012, Smialowski et al., 2012, Fang and Fang, 2013, Hirose and Noguchi, 2013, Agostini et al., 2014). Here we report a novel approach based on protein structure, discounting buried amino acids inside the protein. rHuEPO WT contains a large positively-charged patch on surface (Fig. 4.1), leading to an insoluble prediction according to our calculations (Table 4.1). This finding correlates with previous experimental work (Cheetham et al., 1998, Narhi et al., 2001). Starting from this correlation, we generated a small set of residue candidates to be mutated *in silico* (Supplementary Table S4.2). In order to make a more soluble protein, substitutions to aspartic acid were carried out in residues located within the largest positively-charged patch (rHuEPO F48D, R150D and in combination). In contrast, expanding the posQ patch (Fig. 4.1) by changing a negative charge residue to a positive charged amino acid within the patch (rHuEPO E13K) generated a more insoluble variant (Table 4.1). This set of proteins is intended to give a two-way validation of the algorithm. Previous site-directed mutagenesis of these residues had shown no alteration of the folded state of rHuEPO structure (Elliott et al., 1997). Although the multiple alignment of 50 homologues of HuEPO showed a degree of conservation for R150 (Fig. 4.3), we decided to mutate it based on the surface patch analysis. The evolutionary distribution of positive charge patches among 115 EPO orthologues reveals that HuEPO is located roughly in the centre of the distribution (Fig. 4.2), suggesting that EPOs from different species could have varying solubility properties in expression systems, if positive patches are relevant.

Under low induction conditions, we expected a low translation rate and better folding resulting in less IBs formation. Fig. 4.4 shows a constant rHuEPO WT solubility in both *E. coli* strains, ranging around 55%. An engineered larger positively-charged patch successfully increased IBs formation in rHuEPO E13K as predicted. In contrast, we observed a significant increase in solubility for the less posQ variants in the SHuffle strain and rHuEPO R150D in BL21 (DE3) pLysS. Our hypothesis for this effect of positive charge patches is that they may interact with the population of nucleic acids (e.g. mRNAs) in expression systems, on a pathway to (partial) unfolding and aggregation (Chan et al., 2013). However this is speculative and other factors may be at play. For example, there is literature evidence of hydrophilic properties of negatively charged residues being more favourable than positively charged amino acids on the surface for protein solubility (Trevino et al., 2007, Trevino et al., 2008). Some groups have suggested that a difference in contribution of polar residues arises, where carboxyl groups of negatively-charged amino acids (Asp and Glu) bind to water more strongly than do amino and guanidine groups of positively-charged residues (Arg and Lys) (Kuntz, 1971, Collins and Washabaugh, 1985, Collins, 1997, Chong and Ham, 2014). Whatever the molecular mechanism by which the engineering of positively-charged patches is functioning, it appears possible that it is associated with the native, or a near native state. This is since the pattern of predictions are matched precisely for expression in the cytoplasm of the SHuffle system, which favours a folded recombinant protein with correct formation of the two disulphide bonds. In contrast the pattern of predictions is not matched in the BL21 (DE3) pLysS strain, even with a statistical significant decrease in solubility for rHuEPO F48D. It is not clear what is happening in this case, but examination of the EPO structure indicates that substitution of F48 may actually expose more non-polar surface than it removes. In the wild-type EPO, F48 covers V46 and L155, so that the mutation to aspartic acid may expose these hydrophobic residues. The

plysS strain may be more susceptible to this exposure than the SHuffle strain, since it would be less able to refold partially unfolded protein.

A molecular weight shift from 20.4 kDa to ~20.9 kDa is evident for this mutation for expression in both strains (Fig. 4.4A-B). The shift by a single point mutation, could be due to an alteration in the protein's charge or possibly to phosphorylation on the aspartic acid residue (Race et al., 2007). It has been reported that aspartic acid mutation may mimic the effects of phosphorylation leading to a similar SDS-PAGE gel molecular mass shift (Race et al., 2007). In addition, another reason could be the possibility of the flanking tyrosine (Y49) phosphorylation by modifying its configuration.

This Chapter has tested the hypothesis that positive charge patches can influence protein solubility in expression, using the example of HuEPO as a protein that demonstrates only moderate solubility at high level induction. Whilst not all the results in the two expression systems match predictions, it is interesting that all predictions are matched for expression in the SHuffle system, in which correctly folded, native state proteins will, in principle, be favoured. Thus the SHuffle system may provide a better test of features that are associated with the native state (positive charge patches). The results with rHuEPO, for positive charge patch engineered larger and smaller, are therefore encouraging.

In the SHuffle system, we achieved more soluble variants by diminishing the largest positively-charged patch on rHuEPO and, in contrast, by increasing the patch we generated a less soluble mutant. The work sets the stage for further investigation of the implications of positive charge and other features in solubility of protein during expression, looking also at the mechanism by which the features are exerting their influence. Ultimately this approach will contribute to improved understanding and tools for the design of proteins, with enhanced solubility and stability.

4.5 Supplementary data

Residue			Ratio																
ALA	1	POS	1.492	LEU	35	NEG		LEU	69	NEG		ARG	103	POS	0.477	THR	137	POS	1.492
PRO	2	POS	1.492	ASN	36	HYD		LEU	70	NEG		SER	104	HYD		PHE	138	NEG	
PRO	3	HYD		GLU	37	HYD		SER	71	HYD		LEU	105	POS	0.477	ARG	139	POS	1.492
ARG	4	POS	1.492	ASN	38	HYD		GLU	72	HYD		THR	106	POS	0.477	LYS	140	POS	1.492
LEU	5	POS	1.492	ILE	39	NEG		ALA	73	NEG		THR	107	POS	0.477	LEU	141	POS	1.492
ILE	6	POS	0.003	THR	40	NEG		VAL	74	HYD		LEU	108	HYD		PHE	142	POS	1.492
CYS	7	POS	1.492	VAL	41	HYD		LEU	75	HYD		LEU	109	HYD		ARG	143	HYD	
ASP	8	HYD		PRO	42	POS	0.022	ARG	76	HYD		ARG	110	POS	0.014	VAL	144	POS	1.492
SER	9	POS	1.492	ASP	43	HYD		GLY	77	NEG		ALA	111	HYD		TYR	145	HYD	
ARG	10	HYD		THR	44	POS	1.492	GLN	78	NEG		LEU	112	NEG		SER	146	HYD	
VAL	11	HYD		LYS	45	POS	1.492	ALA	79	HYD		GLY	113	NEG		ASN	147	POS	1.492
LEU	12	HYD		VAL	46	HYD		LEU	80	NEG		ALA	114	NEG		PHE	148	POS	1.492
GLU	13	POS	1.492	ASN	47	POS	1.492	LEU	81	NEG		GLN	115	POS	0.014	LEU	149	HYD	
ARG	14	HYD		PHE	48	POS	1.492	VAL	82	NEG		LYS	116	HYD		ARG	150	POS	1.492
TYR	15	NEG		TYR	49	POS	1.492	ASN	83	HYD		GLU	117	HYD		GLY	151	POS	1.492
LEU	16	HYD		ALA	50	HYD		SER	84	HYD		ALA	118	NEG		LYS	152	POS	1.492
LEU	17	NEG		TRP	51	HYD		SER	85	NEG		ILE	119	NEG		LEU	153	HYD	
GLU	18	HYD		LYS	52	NEG		GLN	86	NEG		SER	120	NEG		LYS	154	POS	1.492
ALA	19	POS	0.477	ARG	53	HYD		PRO	87	HYD		PRO	121	NEG		LEU	155	HYD	
LYS	20	POS	1.492	MET	54	HYD		TRP	88	POS	1.492	PRO	122	NEG		TYR	156	HYD	
GLU	21	NEG		GLU	55	HYD		GLU	89	NEG		ASP	123	NEG		THR	157	POS	1.492
ALA	22	POS	0.477	VAL	56	NEG		PRO	90	HYD		ALA	124	HYD		GLY	158	POS	1.492
GLU	23	POS	1.492	GLY	57	NEG		LEU	91	HYD		ALA	125	NEG		GLU	159	POS	1.492
ASN	24	HYD		GLN	58	NEG		GLN	92	HYD		SER	126	HYD		ALA	160	NEG	
ILE	25	HYD		GLN	59	NEG		LEU	93	HYD		ALA	127	HYD		CYS	161	POS	1.492
THR	26	HYD		ALA	60	NEG		HIS	94	POS	0.477	ALA	128	HYD		ARG	162	POS	1.492
THR	27	POS	1.492	VAL	61	HYD		VAL	95	POS	0.477	PRO	129	HYD		THR	163	POS	1.492
GLY	28	HYD		GLU	62	HYD		ASP	96	NEG		LEU	130	POS	0.025	GLY	164	POS	1.492
CYS	29	POS	1.492	VAL	63	HYD		LYS	97	HYD		ARG	131	HYD		ASP	165	NEG	
ALA	30	POS	1.492	TRP	64	HYD		ALA	98	POS	0.477	THR	132	POS	0.022	ARG	166	NEG	
GLU	31	HYD		GLN	65	HYD		VAL	99	HYD		ILE	133	POS	1.492				
HIS	32	NEG		GLY	66	HYD		SER	100	HYD		THR	134	POS	1.492				
CYS	33	NEG		LEU	67	HYD		GLY	101	NEG		ALA	135	POS	1.492				
SER	34	POS	1.492	ALA	68	HYD		LEU	102	HYD		ASP	136	POS	1.492				

Table S4.1. Positively-charged patches size profile of rHuEPO WT from the charged patch calculator. Complete screening of posQ ratio scores for the modified rHuEPO WT (PDB ID: 1EER) is shown. The largest positive patches are represented by blue (ratio > 1.0). Those proteins with ratio above 1.0 are predicted as insoluble and below 1.0 as soluble. The three targeted residues in this study are highlighted in red. Ratio: largest positively-charged patch (posQ) value from the charged patch calculator (Chan et al., 2013). Charge patches: HYD, hydrophobic (non-charged); NEG, negatively-charged; POS, positively-charged.

Substitution of K-R in the posQ for D		Substitution of residues in the posQ for D				Substitution of D-E in the posQ for K-R	
rHuEPO	posQ	rHuEPO	posQ	rHuEPO	posQ	rHuEPO	posQ
WT	1.49	WT	1.49	I133D	1.40	E13K	2.471
R4D	1.29	S9D	0.89	A135D	1.39	E13R	2.222
K20D	1.23	T27D	1.39	L141D	1.40	E23R	1.63
K45D	0.68	A30D	1.42	F142D	1.39	D136R	1.71
R139D	1.28	S34D	1.42	V144D	1.38	E159R	1.79
K140D	1.26	T44D	1.35	N147D	1.19		
R150D	0.61	N47D	1.36	F148D	1.35		
K152D	0.75	F48D	0.75	G151D	0.74		
K154D	0.77	Y49D	1.43	T157D	0.87		
R162D	0.85	W88D	1.42	G158D	0.84		

Table S4.2. Summary of the solubility screening of rHuEPO. Left column shows the complete mutational screening of all positive charge amino acids (i.e. arginine and lysine) within the largest positively-charged patch (posQ) for aspartic acid (D). Next two columns summarise a set of substitutions of any amino acid in the posQ for D. The column on the right shows all the negative charge residues (i.e. aspartic and glutamic acid) within the posQ for arginine or lysine. Those proteins with posQ ratio above 1.0 are predicted as insoluble and below 1.0 as soluble. Selected proteins for further site-directed mutagenesis are highlighted in red.

Acknowledgements

We would like to thank Dr. E. McKenzie, Dr. E. Ceh and Dr. J. Suarez for supply of the *Escherichia coli* strains and expression vectors, and also CONACyT for contributing PhD funds.

4.6 References

AGOSTINI, F., CIRILLO, D., LIVI, C. M., DELLI PONTI, R. & TARTAGLIA, G. G. 2014. ccSOL omics: a webserver for solubility prediction of endogenous and heterologous expression in Escherichia coli. *Bioinformatics,* 30**,** 2975-7.

AHN, J. H., KEUM, J. W. & KIM, D. M. 2011. Expression screening of fusion partners from an E. coli genome for soluble expression of recombinant proteins in a cell-free protein synthesis system. *PLoS One,* 6**,** e26875.

ASHKENAZY, H., EREZ, E., MARTZ, E., PUPKO, T. & BEN-TAL, N. 2010. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res,* 38**,** W529-33.

BANKS, D. D. 2011. The Effect of Glycosylation on the Folding Kinetics of Erythropoietin. *Journal of Molecular Biology,* 412**,** 536-550.

BUCHANAN, A., FERRARO, F., RUST, S., SRIDHARAN, S., FRANKS, R., DEAN, G., MCCOURT, M., JERMUTUS, L. & MINTER, R. 2012. Improved drug-like properties of therapeutic proteins by directed evolution. *Protein Eng Des Sel,* 25**,** 631-8.

CELNIKER, G., NIMROD, G., ASHKENAZY, H., GLASER, F., MARTZ, E., MAYROSE, I., PUPKO, T. & BEN-TAL, N. 2013. ConSurf: Using Evolutionary Data to Raise Testable Hypotheses about Protein Function. *Israel Journal of Chemistry,* 53**,** 199-206.

CHAN, P., CURTIS, R. A. & WARWICKER, J. 2013. Soluble expression of proteins correlates with a lack of positively-charged surface. *Sci Rep,* 3**,** 3333.

CHEETHAM, J. C., SMITH, D. M., AOKI, K. H., STEVENSON, J. L., HOEFFEL, T. J., SYED, R. S., EGRIE, J. & HARVEY, T. S. 1998. NMR structure of human erythropoietin and a comparison with its receptor bound conformation. *Nat Struct Biol,* 5**,** 861-6.

CHESSHYRE, J. & HIPKISS, A. 1989. Low temperatures stabilize interferon α-2 against proteolysis in Methylophilus methylotrophus and Escherichia coli. *Applied Microbiology and Biotechnology,* 31**,** 158-162.

CHONG, S. H. & HAM, S. 2014. Interaction with the surrounding water plays a key role in determining the aggregation propensity of proteins. *Angew Chem Int Ed Engl,* 53**,** 3961-4.

COLLINS, K. D. 1997. Charge density-dependent strength of hydration and biological structure. *Biophys J,* 72**,** 65-76.

COLLINS, K. D. & WASHABAUGH, M. W. 1985. The Hofmeister effect and the behaviour of water at interfaces. *Q Rev Biophys,* 18**,** 323-422.

DAS, D. & GEORGIADIS, M. M. 2001. A directed approach to improving the solubility of Moloney murine leukemia virus reverse transcriptase. *Protein Sci,* 10**,** 1936-41.

DAVIS, G. D., ELISEE, C., NEWHAM, D. M. & HARRISON, R. G. 1999. New fusion protein systems designed to give soluble expression in Escherichia coli. *Biotechnol Bioeng,* 65**,** 382-8.

DAVIS, J. M., ARAKAWA, T., STRICKLAND, T. W. & YPHANTIS, D. A. 1987. Characterization of recombinant human erythropoietin produced in Chinese hamster ovary cells. *Biochemistry,* 26**,** 2633-8.

ELLIOTT, S., LORENZINI, T., CHANG, D., BARZILAY, J. & DELORME, E. 1997. Mapping of the active site of recombinant human erythropoietin. *Blood,* 89**,** 493-502.

FAN, D., LI, Q., KORANDO, L., JEROME, W. G. & WANG, J. 2004. A monomeric human apolipoprotein E carboxyl-terminal domain. *Biochemistry,* 43**,** 5055-64.

FANG, Y. & FANG, J. 2013. Discrimination of soluble and aggregation-prone proteins based on sequence information. *Mol Biosyst,* 9**,** 806-11.

FAREWELL, A. & NEIDHARDT, F. C. 1998. Effect of temperature on in vivo protein synthetic capacity in Escherichia coli. *J Bacteriol,* 180**,** 4704-10.

FERRER, M., CHERNIKOVA, T. N., YAKIMOV, M. M., GOLYSHIN, P. N. & TIMMIS, K. N. 2003. Chaperonins govern growth of Escherichia coli at low temperatures. *Nat Biotechnol,* 21**,** 1266-7.

FINK, A. L. 1998. Protein aggregation: folding aggregates, inclusion bodies and amyloid. *Fold Des,* 3**,** R9-23.

GUEX, N. & PEITSCH, M. C. 1997. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis,* 18**,** 2714-23.

HIROSE, S. & NOGUCHI, T. 2013. ESPRESSO: a system for estimating protein expression and solubility in protein expression systems. *Proteomics,* 13**,** 1444-56.

HUANG, H. L., CHAROENKWAN, P., KAO, T. F., LEE, H. C., CHANG, F. L., HUANG, W. L., HO, S. J., SHU, L. S., CHEN, W. L. & HO, S. Y. 2012. Prediction and analysis of protein solubility using a novel scoring card method with dipeptide composition. *BMC Bioinformatics,* 13 Suppl 17**,** S3.

JENKINS, T. M., HICKMAN, A. B., DYDA, F., GHIRLANDO, R., DAVIES, D. R. & CRAIGIE, R. 1995. Catalytic domain of human immunodeficiency virus type 1 integrase: identification of a soluble mutant by systematic replacement of hydrophobic residues. *Proc Natl Acad Sci U S A,* 92**,** 6057-61.

JEONG, T. H., SON, Y. J., RYU, H. B., KOO, B. K., JEONG, S. M., HOANG, P., DO, B. H., SONG, J. A., CHONG, S. H., ROBINSON, R. C. & CHOE, H. 2014. Soluble expression and partial purification of recombinant human erythropoietin from E. coli. *Protein Expr Purif,* 95**,** 211-8.

KRANTZ, S. B. 1991. Erythropoietin. *Blood,* 77**,** 419-34.

KUNTZ, I. D. 1971. Hydration of macromolecules. III. Hydration of polypeptides. *Journal of the American Chemical Society,* 93**,** 514-516.

LAI, P. H., EVERETT, R., WANG, F. F., ARAKAWA, T. & GOLDWASSER, E. 1986. Structural characterization of human erythropoietin. *J Biol Chem,* 261**,** 3116-21.

LAWSON, A. J., WALKER, E. A., WHITE, S. A., DAFFORN, T. R., STEWART, P. M. & RIDE, J. P. 2009. Mutations of key hydrophobic surface residues of 11 beta-hydroxysteroid dehydrogenase type 1 increase solubility and monodispersity in a bacterial expression system. *Protein Sci,* 18**,** 1552-63.

LI, Y., YAN, Y., ZUGAY-MURPHY, J., XU, B., COLE, J. L., WITMER, M., FELOCK, P., WOLFE, A., HAZUDA, D., SARDANA, M. K., CHEN, Z., KUO, L. C. & SARDANA, V. V. 1999. Purification, solution properties and crystallization of SIV integrase containing a continuous core and C-terminal domain. *Acta Crystallogr D Biol Crystallogr,* 55**,** 1906-10.

LOBSTEIN, J., EMRICH, C., JEANS, C., FAULKNER, M., RIGGS, P. & BERKMEN, M. 2012. SHuffle, a novel Escherichia coli protein expression strain capable of correctly folding disulfide bonded proteins in its cytoplasm. *Microbial Cell Factories,* 11**,** 56.

MAGNAN, C. N., RANDALL, A. & BALDI, P. 2009. SOLpro: accurate sequence-based prediction of protein solubility. *Bioinformatics,* 25**,** 2200-7.

NARHI, L. O., ARAKAWA, T., AOKI, K., WEN, J., ELLIOTT, S., BOONE, T. & CHEETHAM, J. 2001. Asn to Lys mutations at three sites which are N-glycosylated in the mammalian protein decrease the aggregation of Escherichia coli-derived erythropoietin. *Protein Eng,* 14**,** 135-40.

NARHI, L. O., ARAKAWA, T., AOKI, K. H., ELMORE, R., ROHDE, M. F., BOONE, T. & STRICKLAND, T. W. 1991. The effect of carbohydrate on the structure and stability of erythropoietin. *J Biol Chem,* 266**,** 23022-6.

NIU, X. H., HU, X. H., SHI, F. & XIA, J. B. 2012. Predicting protein solubility by the general form of Chou's pseudo amino acid composition: approached from chaos game representation and fractal dimension. *Protein Pept Lett,* 19**,** 940-8.

NIWA, T., YING, B. W., SAITO, K., JIN, W., TAKADA, S., UEDA, T. & TAGUCHI, H. 2009. Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of Escherichia coli proteins. *Proc Natl Acad Sci U S A,* 106**,** 4201-6.

PERCHIACCA, J. M., LADIWALA, A. R., BHATTACHARYA, M. & TESSIER, P. M. 2012. Aggregation-resistant domain antibodies engineered with charged mutations near the edges of the complementarity-determining regions. *Protein Eng Des Sel,* 25**,** 591-601.

PRICE, W. N., 2ND, HANDELMAN, S. K., EVERETT, J. K., TONG, S. N., BRACIC, A., LUFF, J. D., NAUMOV, V., ACTON, T., MANOR, P., XIAO, R., ROST, B., MONTELIONE, G. T. & HUNT, J. F. 2011. Large-scale experimental studies show unexpected amino acid effects on protein expression and solubility in vivo in E. coli. *Microb Inform Exp,* 1**,** 6.

RACE, P. R., SOLOVYOVA, A. S. & BANFIELD, M. J. 2007. Conformation of the EPEC Tir protein in solution: investigating the impact of serine phosphorylation at positions 434/463. *Biophys J,* 93**,** 586-96.

SACKMANN, E. 1995. Chapter 1 Biological membranes architecture and function. *In:* LIPOWSKY, R. & SACKMANN, E. (eds.) *Handbook of Biological Physics.* North-Holland.

SAMAK, T., GUNTER, D. & WANG, Z. 2012. Prediction of protein solubility in E. coli. *Proceedings of the 2012 IEEE 8th International Conference on E-Science (e-Science).* IEEE Computer Society.

SCHRÖDINGER, L. L. C. 2010. The PyMOL Molecular Graphics System, Version 1.3r1.

SEVASTSYANOVICH, Y., ALFASI, S., OVERTON, T., HALL, R., JONES, J., HEWITT, C. & COLE, J. 2009. Exploitation of GFP fusion proteins and stress avoidance as a generic strategy for the production of high-quality recombinant proteins. *FEMS Microbiol Lett,* 299**,** 86-94.

SLOVIC, A. M., SUMMA, C. M., LEAR, J. D. & DEGRADO, W. F. 2003. Computational design of a water-soluble analog of phospholamban. *Protein Sci,* 12**,** 337-48.

SMIALOWSKI, P., DOOSE, G., TORKLER, P., KAUFMANN, S. & FRISHMAN, D. 2012. PROSO II--a new method for protein solubility prediction. *FEBS J,* 279**,** 2192-200.

SMIALOWSKI, P., MARTIN-GALIANO, A. J., MIKOLAJKA, A., GIRSCHICK, T., HOLAK, T. A. & FRISHMAN, D. 2007. Protein solubility: sequence based prediction and experimental verification. *Bioinformatics,* 23**,** 2536-42.

SUZEK, B. E., HUANG, H., MCGARVEY, P., MAZUMDER, R. & WU, C. H. 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics,* 23**,** 1282-1288.

TREVINO, S. R., SCHOLTZ, J. M. & PACE, C. N. 2007. Amino acid contribution to protein solubility: Asp, Glu, and Ser contribute more favorably than the other hydrophilic amino acids in RNase Sa. *J Mol Biol,* 366**,** 449-60.

TREVINO, S. R., SCHOLTZ, J. M. & PACE, C. N. 2008. Measuring and increasing protein solubility. *J Pharm Sci,* 97**,** 4155-66.

WALSH, G. 2014. Biopharmaceutical benchmarks 2014. *Nat Biotech,* 32**,** 992-1000.

WANG, W. 2005. Protein aggregation and its inhibition in biopharmaceutics. *Int J Pharm,* 289**,** 1-30.

WILKINSON, D. L. & HARRISON, R. G. 1991. Predicting the solubility of recombinant proteins in Escherichia coli. *Biotechnology (N Y),* 9**,** 443-8.
ZHANG, Y. B., HOWITT, J., MCCORKLE, S., LAWRENCE, P., SPRINGER, K. & FREIMUTH, P. 2004. Protein aggregation during overexpression limited by peptide extensions with large net negative charge. *Protein Expr Purif,* 36**,** 207-16.

Chapter 5

Paper 3:

Modulation of recombinant erythropoietin secretion in HEK 293-EBNA cells through modification of protein surface patches

Carballo-Amador M.A.[12], Warwicker J.[2] and Dickson A.J.[1]

[1] Faculty of Life Sciences, University of Manchester, Michael Smith Building, Oxford Road, Manchester M13 9PT, UK

[2] Faculty of Life Sciences, University of Manchester, Manchester Institute of Biotechnology, 131 Princess Street, Manchester M1 7DN, UK

Abstract

Protein-based drugs form the basis for the biopharmaceutical manufacturing industry. Currently, more than half of all recombinant therapeutic proteins have been produced in mammalian cells, mainly due to the high similarity of the final product to human protein structures. Most of the commercial therapeutics are secreted proteins, such as hormones and monoclonal antibodies. Poor secretion may be due to misfolded proteins or aggregates leading to cellular stress and proteolysis. Therefore, maximising the secretory capacity is a potential area to exploit. We have previously investigated the solubility of a synthetic redesigned set of recombinant human erythropoietin (rHuEPO) variants with altered surface charge upon heterologous expression in *E. coli*. Here we report the consequences for production of this set of rHuEPO variants when transiently expressed in human embryonic kidney 293 EBNA (HEK 293-EBNA) cell line. We found that the variant predicted to be less soluble (rHuEPO E13K) decreased secretion by half compared to the rHuEPO WT. In contrast, those predicted to be soluble (i.e. rHuEPO F48D, R150D, and F48D/R150D) increased by 2 to 4-fold secretion compared to rHuEPO at day five of culture. Our findings suggested that positively-charged patch size may influence protein secretion.

5.1 Introduction

In the field of recombinant protein technology, cost-efficient production of protein of necessary quality and quantity is the pinnacle goal for manufacture of biopharmaceuticals. However, expression of a heterologous gene in diverse cellular environments is associated with several challenges. These range from low yield production (e.g. upstream and downstream process) to undesirable protein modifications (e.g. protein aggregation leading to immunogenicity in clinical use) (Wang et al., 2012, Ratanji et al., 2014). To address these issues several strategies have been developed, which include modification of environmental conditions, such as temperature and nutrient balance, genetic and cellular engineering (e.g. codon optimisation, vector design, trafficking) and protein engineering. This latter strategy seeks to improve protein physicochemical properties and hence enhance expressibility (Derewenda, 2010, Close et al., 2014). Although proteins have achieved their natural structure (and function) through millions of years of evolutionary pressure, their properties can still be improved or altered (Jäckel et al., 2008).

Human erythropoietin (HuEPO) presents an interesting exemplar, especially in relation to the high economic expectations for biosimilar products and analogues (biobetters) (Jelkmann, 2013, Mikhail and Farouk, 2013). HuEPO consists of an initial immature protein of 193 amino acids, which includes a 27 amino acid signal peptide that is removed along with the carboxyl-terminal arginine in maturation to circulatory format (Jacobs et al., 1985, Recny et al., 1987). HuEPO with a molecular mass ~30.4 kDa, (Davis et al., 1987), contains three potential N-linked (N24, N38 and N83) sites and one potential O-linked glycosylation site (S126) (Lai et al., 1986) and the glycosylation (that has an essential role in the appropriate bioactivity of HuEPO) can contribute ~40% of the molecular mass of the mature protein (Higuchi et al., 1992). Potential immunogenicity remains a major concern in the production of

biosimilar formats of recombinant human erythropoietin (rHuEPO) (McKoy et al., 2008, Brinks et al., 2011, Praditpornsilpa et al., 2011). This may be caused by its heterogeneous glycosylation (Lisowska, 2002, Kamioner, 2012) or by formation of high molecular aggregates (Park et al., 2009, Brinks et al., 2011). Protein aggregation of therapeutics is an undesirable phenomenon at any stage of production or clinical administration (Wang et al., 2012), which may arise from interactions of proteins with incorrect folding *in vivo* or physical insults during handling (Wetzel, 1996), interactions with materials involved in formulation (Seidl et al., 2012) or by inappropriate storage (Fotiou et al., 2009).

Buchanan and co-workers have shown (by a random mutagenesis approach followed by ribosome display) that it is possible to engineer improved properties into variants of rHuEPO (Buchanan et al., 2012). A number of the resultant variants were more active and less prone to aggregation than natural (wild type, WT) HuEPO. rHuEPO has also been subjected to rational design via engineering of additional N-linked glycosylation sites (Elliott et al., 2003, Su et al., 2010). Previous studies have shown that alterations to rHuEPO glycosylation influenced half-life in serum (Elliott et al., 2003), stability under denaturing conditions (e.g. pH, temperature) *in vitro* (Narhi et al., 1991) and rates of clearance (e.g. sialylation of the terminal sugar in N-linked glycan) *in vivo* (Jeong et al., 2008).

We have previously reported a novel approach to predict protein solubility based on surface charge patches and consequently engineered rHuEPO for improved solubility when expressed in *E. coli* (Carballo-Amador et al., 2014b). This approach emerged from a recently published hypothesis from our group, stating that soluble expression of proteins inversely correlated with the size of the largest positively-charged patch on a protein surface (Chan et al., 2013). Five rHuEPO constructs (rHuEPO WT, rHuEPO E13K, rHuEPO F48D, rHuEPO R150D and rHuEPO F48D/R150D) with different surface profile charge were described in a previous study (Carballo-Amador et al., 2014b). Those variants with diminished positively-

charged patches (i.e. mutations from positive to negative charge residue) resulted in more soluble proteins in expression.

In the present paper, we examined the consequences for production of this set of rHuEPO variants when transiently expressed in human embryonic kidney 293 EBNA (HEK 293-EBNA) cell line. HEK-293 cells are widely used as a transient expression system due to high transfection efficiency and their capacity to support protein production (Thomas and Smart, 2005). In addition, this cell line is often used in preliminary screening of potential proteins for stable or large-scale production. Here we report correlation between poor product yield and the extent of positive charge on protein surface patches. Further application of this approach could offer a tool for the rational design of proteins for therapeutic and other uses.

5.2 Material and methods

5.2.1 *Computational calculations*

An algorithm (Chan et al., 2013) was used to compute surface charged of rHuEPO WT and variants (Carballo-Amador et al., 2014b). This approach has been applied to prediction of the solubility of rHuEPO G09 variant developed by a MedImmune research group at Cambridge (Buchanan et al., 2012) (Table 5.1). The algorithm computes structured-based parameters, including, nonQmax –the maximal non-charged patch-, posQmax –the maximal size of a positively charged patch- and their multiplicative combination, versus thresholds calculated from Niwa dataset of experimental solubilities determined for cell-free expression of *E. coli* proteins (Niwa et al., 2009). These thresholds have been derived by looking for the value of a parameter that best separates soluble proteins, in terms of their extent of solubility, with those

proteins with the largest positive patch predicted as least soluble (ratio above 1.0 relative to the threshold). Positive patches with ratio below 1.0 are predicted as more soluble. In order to generate a rHuEPO G09 structure (I25F, T27S, R139H and G158E), substitutions on the crystal structure of rHuEPO (PDB ID: 1EER) were carried out in Swiss-PdbViewer by applying the "MUTATE" function. Then, new PDB files were generated for further analysis in the Chan et al., (2013) algorithm (Table 5.1). The PyMOL Molecular Graphics System version 1.3 (Schrödinger, 2010) was used to render protein surface structures. Theoretical isoelectric point (pI) was calculated using ProtParam tool (Gasteiger et al., 2005).

5.2.2 *Construction of mammalian expression vectors*

Human erythropoietin cDNA was amplified from a pre-existing mammalian expression vector (pCET-901-HuEPO, Section 9.3) by utilisation of primers containing the addition of the restriction site 5'-*HindIII* and 3'-6x His-tag (For 5' CTAAAGCTTATGGGGGTGCAC GAATGTCCTG 3'; Rev 5' GATCAGTGGTGGTGGTGATGGTGTCTGTCCCCTGTCCT GCAGGCCTC 3'). An intermediary subcloning was carried out into a TA cloning vector (pCR2.1, Invitrogen). This product was subjected to a second PCR to add the restriction site *NotI* at 3'of the target template (For 5' CTAAAGCTTATGGGGGTGCAC GAATGTCCTG 3'; Rev 5' GTAGCGGCCGCTCAGTGGTG GTGGTGATGGTG 3'). The PCR products were restricted with *HindIII* and *NotI* and cloned into the pCEP-PU vectors (kindly provided by Dr. Edward McKenzie of the University of Manchester) under the control of the human cytomegalovirus (CMV) promoter. The resulting sequence for each vector was as follows: 5'-*HindIII*-Signal Peptide-rHuEPO-6xHis-*NotI*-3'. Cloning and propagation of all the vectors were performed in chemically-competent *Escherichia coli* XL1-Blue cells (Stratagene). The host cells were grown in Luria-Bertani (LB) Broth (10g tryptone, 5g yeast extract, 5g NaCl)

containing 100µg/ml ampicillin for plasmid maintenance at 37ºC with shaking at 220 rpm. Small and large scale purifications of plasmid DNA were carried out by QIAGEN Plasmid Mini and Midi Kits, respectively.

5.2.3 *HEK 293-EBNA culture and transient protein expression*

The HEK 293-EBNA cells were kindly provided by Dr. Edward McKenzie from the protein expression facility at the University of Manchester. This cell line stably expresses the EBNA-1 gene (Epstein-Barr virus nuclear antigen 1), allowing episomal replication of the pCEP-PU plasmid. This interaction is essential for DNA replication and distribution of the vector in daughter cells ensuring high protein expression. HEK 293-EBNA cells were grown as monolayers in a humidified incubator with 5% CO_2 at 37ºC. This adherent cell line was cultured in growth medium (Dulbecco's Modified Eagle's Medium [DMEM] [Sigma-Aldrich, # D5796] supplemented with 2 mM L-Glutamine [Sigma-Aldrich, # G7513], 10% [v/v] filtered Foetal Bovine Serum [Life Technologies, #10500-064]) with or without 250 µg/ml Geneticin (G418 Sulphate, Life Technologies, #10131-027) in T75 flasks with filter caps (Corning Incorporated). A transfection efficiency analysis using a GFP plasmid (pCEP-PU-EmGFP, Section 9.4) was carried out in prior transfections (Supplementary Figure S5.1). To prepare cells for transfection, $5x10^5$ cells/ml were incubated overnight in 2.5 ml growth medium (without Geneticin) in each well of 6-well plates, then (with cells $\geq$80% confluent) 2.5 µg of the appropriate pCEP-PU-rHuEPO plasmid was mixed with 7.5 µl *Trans*IT-LT1 Reagent (Mirus), diluted with addition of 250 µl of Opti-MEM I Reduced-Serum Medium (Life Technologies, #31985-062) in separate sterile tubes and incubated at room temperature for 25 minutes. A mock transfection control was carried out with all the components apart from plasmid DNA. For each transfection, 260 µl of the *Trans*IT-LTI Reagent:DNA complex

mixture was added in a drop-wise fashion to each well. Cells were subsequently cultured in a humidified incubator with 5% CO_2 at 37ºC for up to seven days.

5.2.4 *Determination of cell density and viability*

Total cell and viable cell numbers were determined by light microscope using a Neubauer haemocytometer and a Countess Automated Cell Counter (Life Technologies, Invitrogen) after 1:1 dilution of cell suspension with trypan blue solution (0.5% [w/v] in 1x PBS).

5.2.5 *Protein sample preparation*

Medium samples (100 µl) were collected each day. For intracellular samples, cells were harvested at days 5 and 7 of culture by trypsinization and cells were pelleted by centrifugation (100 g, 5 min, room temperature). The cell pellets were washed in 1x PBS and re-centrifuged before resuspension in 40 µl 1x denaturing Laemmli SDS-PAGE sample buffer (63 mM Tris, pH 6.8, 2% [w/v] SDS, 10% [v/v] glycerol, 0.01% [w/v] bromophenol blue and 5% [v/v] β-mercaptoethanol) per $1x10^6$ cells. Cell disruption was maximised by passage through a 23 gauge needle 10 times. All protein samples were boiled for 5 min in sample buffer and stored at -80ºC prior to analysis.

5.2.6 *SDS-PAGE and Western blot*

The medium and cellular protein samples were resolved by electrophoresis on 12.5% (w/v) sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE) using the Mini-

PROTEAN Tetra Cell (BioRad). Equal volumes of medium samples were heated at 95°C by 5 min in 6x denaturing buffer (375 mM Tris pH 6.8, 12% [w/v] SDS, 60% [v/v] glycerol, 0.06% [w/v] bromophenol blue, 15% [v/v] β-mercaptoethanol). Medium and intracellular ($1x10^6$ cells) samples were separated at room temperature in electrode running buffer (50 mM Tris, 0.38 M glycine, 0.2% [w/v] SDS) at 60 V until samples passed into the separating gel and then the voltage was increased to 160 V. For western blotting, proteins were transferred to nitrocellulose membranes (Pall) (equilibrated with blotting buffer, 25 mM Tris, pH7.4, 0.2 M glycine and 20% [v/v] methanol) after their separation by SDS-PAGE. The transfer was accomplished using a transblot semi-dry transfer cell (Bio-Rad) at 15V for 45 min. After blocking the membrane in blocking buffer (5% [w/v] skimmed milk in PBS-Tween pH 7.4) for 2 h at room temperature with rotation, the membrane was incubated for 12-14 h in blocking buffer solution containing mouse anti-HuEPO monoclonal antibody (R&D Systems; MAB287) (1:1500 dilution) at 4°C in rotation. Next day, the primary antibody was removed and the membrane was washed three times (5 min each time) with PBS-Tween at room temperature. For detection of rHuEPO WT and variants, an IR-labelled secondary Donkey anti-Mouse IgG antibody (LI-COR) (1:15000 dilution) in blocking buffer solution was added and incubated for 45 min at room temperature in rotation. Following the incubation, the secondary antibody was removed and the membrane was washed three times as mentioned before. For IR detection and protein staining, blots and gels were imaged with the Odyssey Imaging System. Bands were quantified in Image Studio Lite software (LI-COR).

For specific detection of ERK 2, membranes containing cell lysate samples were stripped with acidic glycine stripping buffer (0.1 M glycine, 20 mM Magnesium acetate, 50 mM KCl, pH 2.2) for 10 minutes with gentle agitation at room temperature. Membranes were washed three times in PBS-Tween for 5 minutes with gentle agitation after the stripping incubation. Then, membranes were re-blocked and incubated with mouse anti-ERK 2

monoclonal antibody (Santa Cruz Biotechnologies, #81459) diluted in blocking buffer (1:2000 dilution) for 12-14 h. Next day, the membranes were processed as described previously.

5.3 Results

5.3.1 *Positively-charged patches on rHuEPO WT predict solubility*

We have previously calculated the solubility profiles of rHuEPO WT and variants using an in-house algorithm (Carballo-Amador et al., 2014b). Here we extended the calculation to the rHuEPO G09 variant (developed by Buchanan and co-workers) a variant with greater biological activity and lesser tendency to aggregation (Buchanan et al., 2012). Proteins with high posQ positively-charged patches (positive patch ratio above 1.0) are predicted to be less soluble (Table 5.1). Single or multiple point mutations can be predicted to have diminished positively-charged patches relative to rHuEPO WT according to our computational calculations (Table 5.1 and Fig. 5.1). Interestingly, rHuEPO G09 mutations are located in, or in close proximity to, the positively-charged patch with greater posQ value (Supplementary Table S5.1). This large positively-charged patch increased in size in rHuEPO WT and E13K calculations (Fig. 5.1). In contrast, negatively-charged mutations diminished the largest positively-charged patch area. The positive patch engineering was designed for alteration of soluble expression in *E. coli*. Since the steps for transient expression from HEK 293-EBNA cells are quite different from cytoplasmic expression in *E. coli*, it is not clear how the engineered EPO proteins will behave. This Chapter describes their observed behaviour, and provides a discussion of possible underlying molecular factors.

TABLE 5.1. Predicted solubilities of rHuEPO WT and variants.

rHuEPO	Pos patch ratio to threshold	Theoretical pI	Solubility prediction
WT	1.49	8.75	Prone to aggregate
F48D	0.75	8.43	Soluble
R150D	0.61	7.92	Soluble
F48D/R150D	0.47	7.27	Soluble
E13K	2.47	9.16	Prone to aggregate
G09 (I25F, T27S, R139H & G158E)	0.79	7.89	Soluble

Solubility prediction profile was defined as described in Chan et al (2013). Positive (Pos) patch sizes are divided by that best separating soluble and insoluble datasets based on cell-free expression of *E. coli* proteins (Niwa et al., 2009), above 1.0 implies predicted aggregation tendency. Theoretical isoelectric point (pI) was calculated using ProtParam tool (Gasteiger et al., 2005). G09 was taken from the experimental from Buchanan et al., 2012.

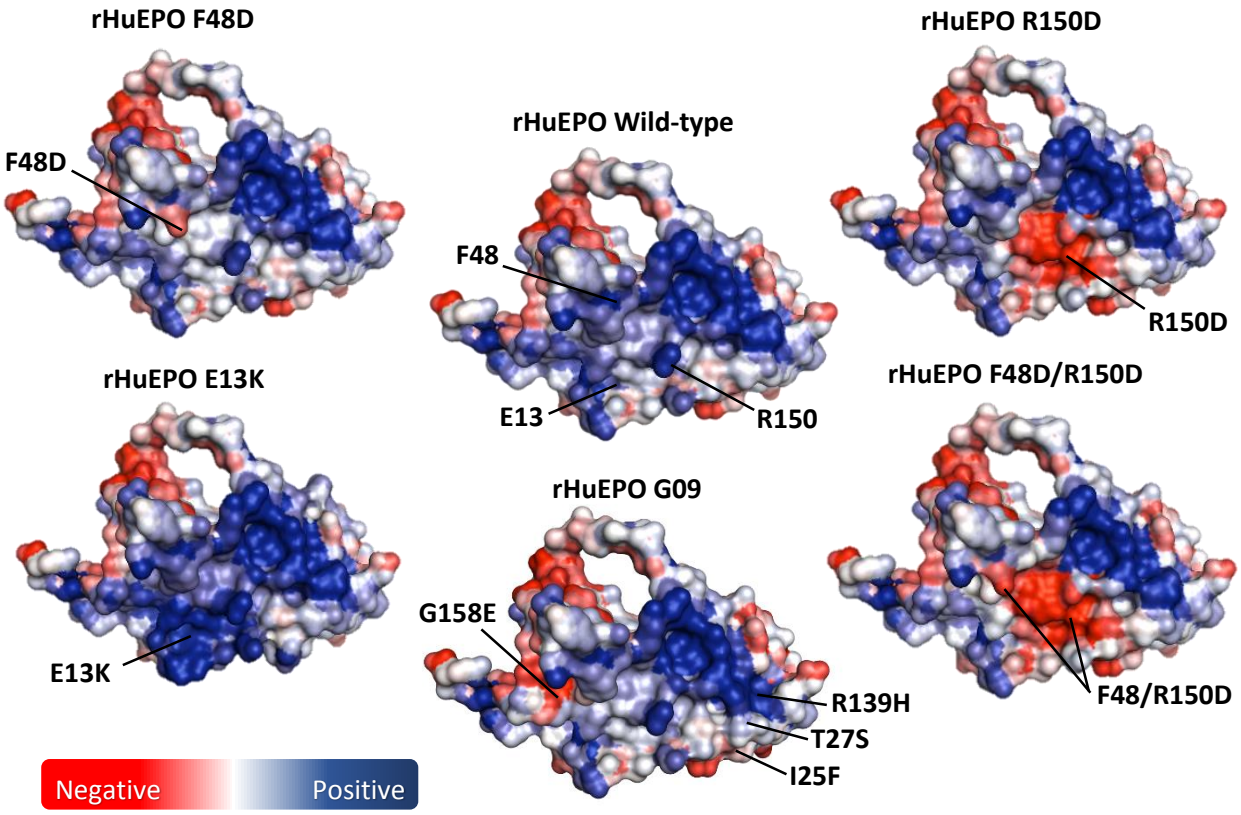


Fig. 5.1. Localisation of target residues on rHuEPO surface. Electrostatic potential computed with the algorithm developed in our group (Chan et al., 2013) onto the molecular surface of rHuEPO and variants. Amino acids in the largest positive patches are represented by blue, non-charged patches by white and negatively charged by red colour, respectively.

5.3.2 *Positively-charged rHuEPO mutations generate decreased rHuEPO secretion*

The substitution of a negatively for a positively charged amino acid (to generate rHuEPO E13K) resulted in an increase in the prediction ratio (compared to the rHuEPO WT, Table 5.1). The computational representation of the amino acid change generated a larger positively-charged surface patch compared to rHuEPO WT (Fig. 5.1). The experimental testing of this hypothesis in an *E. coli* system (Carballo-Amador et al., 2014b) resulted in production of a protein with an increased propensity to aggregate. It is intriguing that the relative propensity of rHuEPO E13K to aggregate in *E. coli* was reflected by the observation that rHuEPO E13K had the lowest secretion of all rHuEPO variants (at both days three and five of culture) (Table 5.2 and Fig. 5.2A-B). Intracellular samples were immunoblotted for rHuEPO (normalised to the loading control ERK 2) at two stages of culture (days five and seven). At day seven some extra bands were visualised compared to day five. These bands ranged from the expected non-glycosylated form (around ~20 kDa) to partially glycosylated forms of ~27 kDa forms (Fig. 5.2C). rHuEPO E13K showed slightly more abundance, especially compared to the WT at both collection days (Table 5.2 and Fig. 5.2D). Cell number of the host HEK cell line was compromised when the rHuEPO E13K variant (with an increased positively-charged patch) was transiently expressed (Fig. 5.3). Cell densities corresponding to samples of day five samples were $9.1x10^6$ cells/ml for rHuEPO E13K and $12.2x10^6$ cells/ml for the control. At day seven of culture cells expressing rHuEPO E13K and control were $9.6x10^6$ cells/ml and $10.8x10^6$ cells/ml, respectively. The growth rate was significantly different for rHuEPO E13K than for non-transfected cells at both days of culture.

TABLE 5.2. rHuEPO WT and variants secretion and intracellular expression profile.

rHuEPO	Aim	Relative total rHuEPO secretion normalised to rHuEPO WT		Relative total intracellular rHuEPO normalised to ERK 2	
		Day 3	Day 5	Day 5	Day 7
WT		1.00 ± 0.17	1.00 ± 0.20	0.88 ± 0.21	2.14 ± 0.42
E13K	Increased posQ	0.57 ± 0.06	0.55 ± 0.04	1.70 ± 0.32	3.12 ± 0.26
F48D	Decreased posQ	3.90 ± 0.30	2.64 ± 0.42	1.65 ± 0.29	4.18 ± 0.30
R150D	Decreased posQ	2.02 ± 0.21	1.19 ± 0.12	1.32 ± 0.32	2.07 ± 0.20
F48D/R150D	Decreased posQ	3.36 ± 0.88	1.63 ± 0.31	1.66 ± 0.31	2.75 ± 0.28

For every HEK 293-EBNA culture, triplicate biological replicates were performed for data generation and deviation represent the ± SEM. posQ, the maximal size of a positively charged patch.

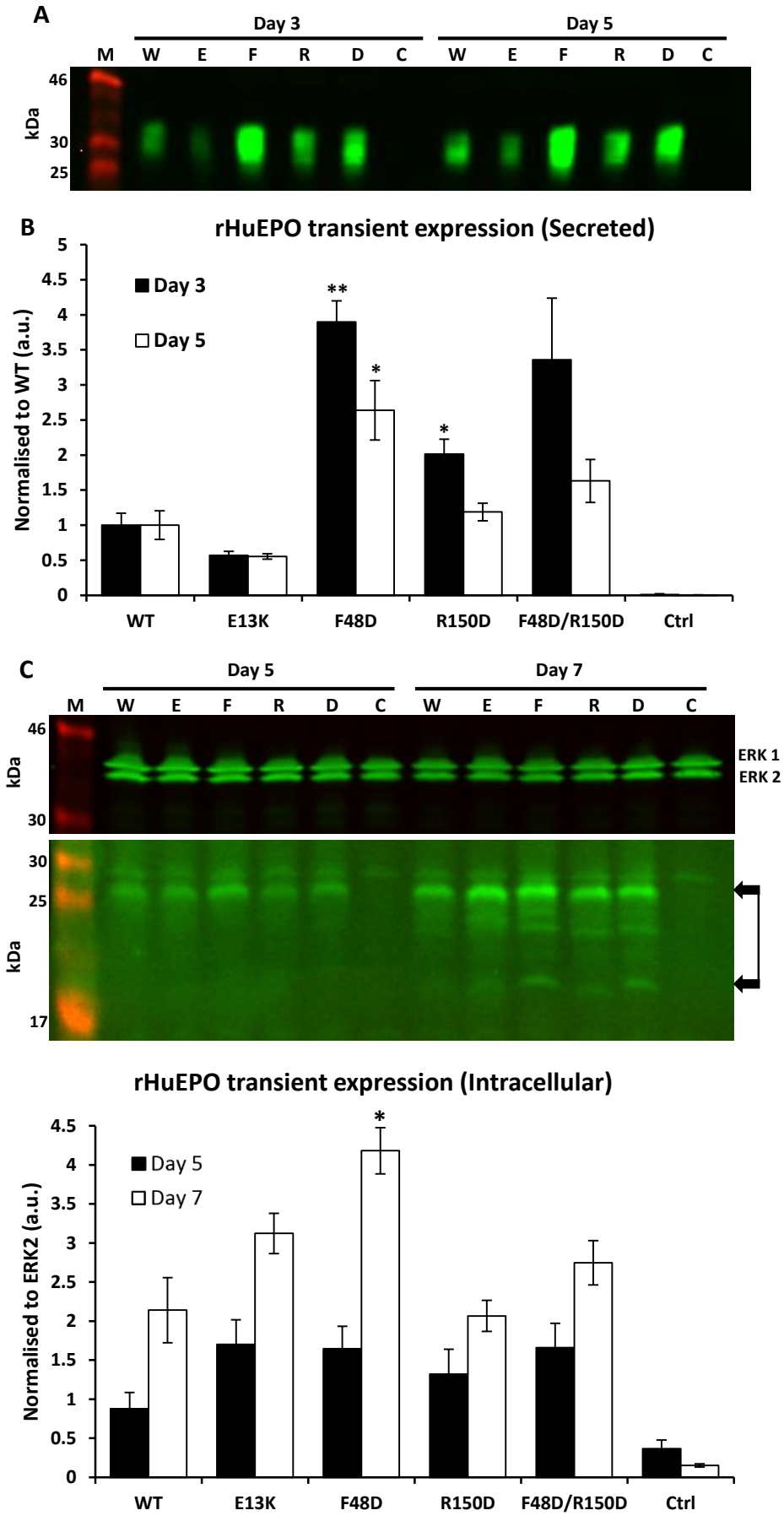

A
Day 3
Day 5
M W E F R D C W E F R D C
kDa
46
30
25
B
rHuEPO transient expression (Secreted)
Normalised to WT (a.u.)
Day 3
Day 5
**
*
*
WT
E13K
F48D
R150D
F48D/R150D
Ctrl
C
Day 5
Day 7
M W E F R D C W E F R D C
kDa
46
30
ERK 1
ERK 2
30
25
17
rHuEPO transient expression (Intracellular)
Normalised to ERK2 (a.u.)
Day 5
Day 7
*
WT
E13K
F48D
R150D
F48D/R150D
Ctrl

Fig. 5.2. Expression analysis of rHuEPO WT and variants. (**A**) Western blot of secreted rHuEPO samples from day three and five into the medium were probed with mouse anti-HuEPO monoclonal antibody and were imaged with the Odyssey Imaging System. (**B**) Relative total rHuEPO secretion normalised to rHuEPO WT from day three and five were plotted as arbitrary units. (**C**) Western blots showing the ERK 2 loading protein control (above) for intracellular rHuEPO expression (below). (**D**) Relative total intracellular rHuEPO production normalised against ERK 2 was plotted as arbitrary units. For every HEK 293-EBNA culture, triplicate biological replicates were performed for data generation and error bars represent the $\pm$ SEM; statistically significant difference was performed using a two-sided unpaired t-test (*P < 0.05, ** P <0.01). M, prestained SDS-PAGE marker (Bio-Rad); W, rHuEPO WT; E, rHuEPO E13K; F, rHuEPO F48D; R, rHuEPO R150D; D, rHuEPO F48D/R150D; C, control (Ctrl).

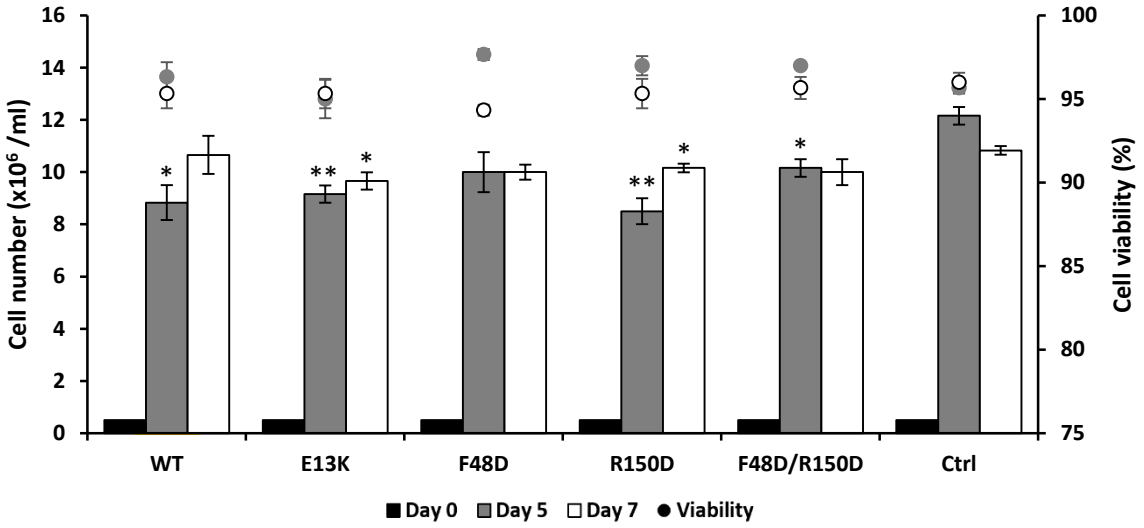


Fig. 5.3. Effect of rHuEPO expression on HEK 293-EBNA cells proliferation. The cell densities and viabilities were determined by trypan blue exclusion assay from day five and seven post-transfection. For every culture, triplicate biological replicates were implemented for data generation and error bars represent the $\pm$ SEM; statistically significant difference was performed using a t-test (*P < 0.05, ** P <0.01).

5.3.3 *Negatively-charged rHuEPO mutations generate increased rHuEPO secretion*

Diffuse bands of the transient expressed rHuEPO were observed with immunoblotting (Fig. 5.2A and C). This is likely to be a reflection of the range of glycosylation heterogeneity of rHuEPO (Skibeli et al., 2001). Immunoreactive bands migrated in the range of ~25–32 kDa corresponding to the estimated molecular mass for rHuEPO (Davis et al., 1987). As quantified in Fig. 5.2B, less secretion was detected for the greatest posQ, positively-charged mutant (rHuEPO E13K) in relation to rHuEPO WT during both days of collection (Day 3 and 5). In contrast, rHuEPO variants containing a negatively-charged mutation (i.e. rHuEPO F48D, R150D and F48D/R150D) revealed more secretion than rHuEPO WT. Statistical significance was achieved for the negatively-charged single mutants at day 3 and only for rHuEPO F48D at day 5 compared to rHuEPO WT. For samples harvested at day five only rHuEPO F48D showed a significant difference from rHuEPO WT.

As mentioned above, cellular extracts analysis showed extra bands, more prominent around the expected non-glycosylated form (~20 kDa) for F48D containing variants (Fig. 5.2C). These bands correlated to a slightly higher molecular weight compared to the WT and the other variants observed in our previous results in *E. coli*-derived non-glycosylated rHuEPO (Carballo-Amador et al., 2014b). At day five, relative intensity of the containing negative charge mutations showed no significant change (Table 5.2 and Fig. 5.2D). In addition, no statistical significant change was observed for samples from day five, although an increase is perceived for all the variants compared to the rHuEPO WT. A statistically significant change was detected only for rHuEPO F48D at late-stage of the cell culture. As shown in Fig. 5.3, cells transiently expressing rHuEPO proliferated slower than control cells at both collection days, mainly at day five of culture.

5.4 Discussion

In this study, we investigated the consequences of expressing rHuEPO variants with altered surface charge in mammalian cells based on a previous work in aggregation profiling upon *E. coli* expression (Carballo-Amador et al., 2014b). We reported a correlation between soluble expression and the computed positively-charge patch size of a rHuEPO set of variants with enhanced (lower positively-charged patch) and decreased solubility (larger positively-charged patch). In this chapter, we found that diminishing the largest positively-charged patch (Table 5.1 and Fig. 5.1) enhanced the recombinant protein yield compared to rHuEPO WT at both sample collection days (day three and five) (Table 5.2 and Fig. 5.2A-B). In contrast, a higher positively-charged patch variant (i.e. rHuEPO E13K) resulted in the lowest amount of secreted protein of all EPO variants, at day three and five. Better secretion may be related to a protein burden on the host cells, a reflection of cell proliferation (Fig. 5.3). In addition, the theoretical isoelectric point among the proteins varies depending on the engineered patch (Table 5.1). Those with largest positively-charged patches tend to have higher pI with converse true for those with least positive charge. The pH gradient decreases along the secretory pathway from 7.2 (endoplasmic reticulum), passing through 6.7 - 6.0 (Golgi) to 5.2 (secretory granulates) (Paroutis et al., 2004). Thus, protein pI could be related to pH variation in the secretory pathway, but it is unclear what the molecular basis of such an effect would be, especially since the EPO forms with the greatest net charge (higher pI) generally secrete less well.

Early work to elucidate HuEPO structure has shown that with one single amino acid substitution is sufficient to alter rHuEPO activity, or in the worst scenario no secretion was observed (Wen et al., 1994). This indicated the importance of a single point mutations in the secretory pathway. In the particular case of the target residues presented here (E13, F48 and R150), no alteration in the folded state was observed by a series of experimental mutations in

rHuEPO (Elliott et al., 1997). In contrast, decreased activity was evident in their results for R150A. We find a correlation between patch calculations and experimental study from a MedImmune research group (Buchanan et al., 2012) on a set of more active mutants. These were obtained from an extensive DNA library by applying random PCR mutagenesis, ribosome display and selection pressures (e.g. reducing agent, elevated temperature and hydrophobic interaction chromatography matrices). In particular, rHuEPO G09 from this library encodes four mutations at low conserved residues (Supplementary Figure S5.2): I25F, T27S, R139H and G158E. Three of these residues were calculated to be inside the largest positively-charged patch, where the exception is the surrounding isoleucine at position 25 (Supplementary Table S5.1). When the mutations were studied individually, the substitution for a negatively-charged amino acid (G158E) decreased the aggregation propensity, an effective not produced by any of the other mutations (I25F, T27S or R139H) (Buchanan et al., 2012). Construct G09 results in a diminished positively-charged patch compared to rHuEPO WT (Table 5.1 and Fig. 5.1), reinforcing the influence of surface charge properties on aggregation propensity.

A parallel study in our group has shown similar ordering of secretion propensity with the same set of EPO variants, in which the proteins lacked of the C-terminal his-tag (Jiun Fu, unpublished data). This study examined, by subcellular fractionation, the distribution of rHuEPO WT and variants along compartments of the secretory pathway. A strong association was observed between those EPO variants with larger positively-charged surface patches and the cytoskeletal fraction. This rationalisation is perhaps the clearest molecular explanation for the observed effects of surface charge engineering in rHuEPO secretion from mammalian cells. It is therefore different to the hypothesis put forward for cytoplasmic over-expression in *E. coli*, which suggests that positively-charged charged protein patches may interact with negatively-charged nucleic acids (Chan et al., 2013). Ubiquitinylation, that takes place primarily on lysine residues of target proteins, may play a part in determination of intracellular

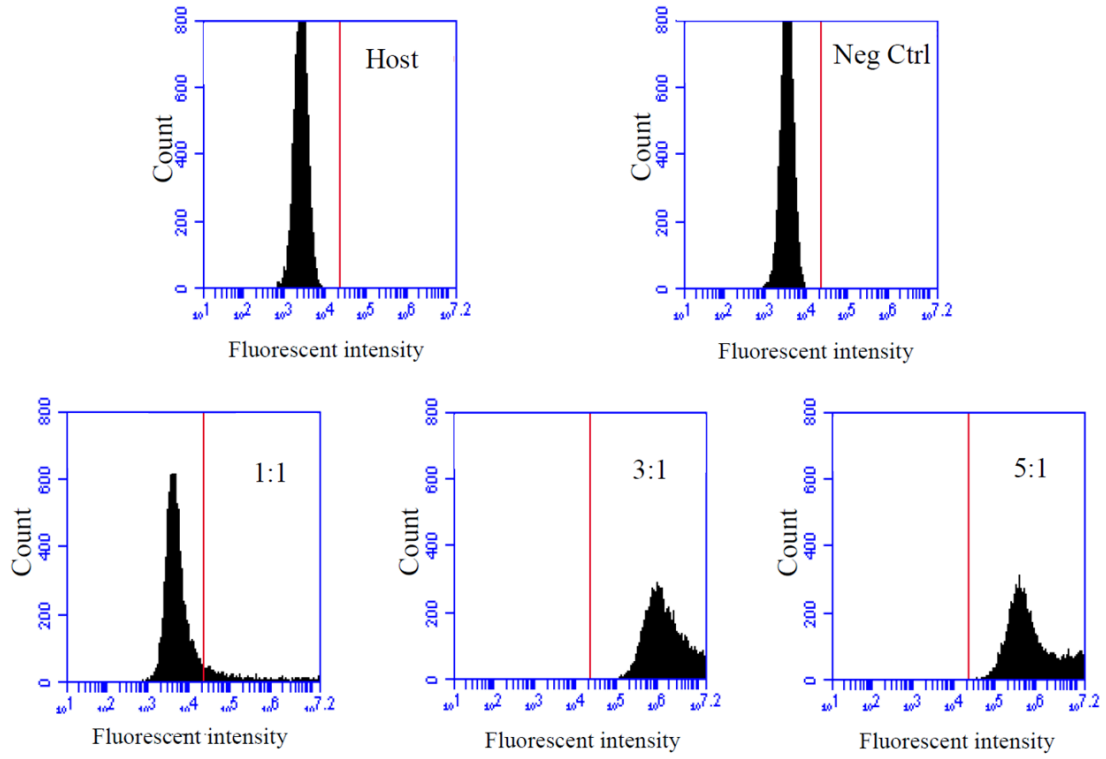
processing of EPO variants. Whilst the exact consensus site for targeted lysine ubiquitylation is poorly understood (Mattiroli and Sixma, 2014), in previous studies of histone H2A two N-terminal lysines (K13 and K15) are targets for ubiquitination (Gatti et al., 2012, Fradet-Turcotte et al., 2013, Mattiroli et al., 2014). Potential ubiquitination of the N-terminal lysine residue on rHuEPO E13K may generate a lysine-targeting modification leading to proteasomal degradation.

To conclude, consequences of changing charged surface amino acids on the surface of rHuEPO were reflected in the amount of secretion. These findings correlate in general terms with previous results in *E. coli* expression system, where larger positively-charged patches led to aggregation differing from the less positively-charged patch variants (Carballo-Amador et al., 2014b). In more detail however, it is not clear that the mechanisms underlying these effects are reproduced between the bacterial and mammalian cell secretion systems. Further examination of protein modifications and cellular responses shall contribute to a better understanding of the consequences of this computational-experimental analysis for the secreting and recovery of engineered recombinant proteins. We hope that our approach will be useful as a starting point to improve recombinant proteins for research and clinical purposes.

5.5 Supplementary data

5.5.1 *Optimising conditions for transfection of HEK 293-EBNA cells*

A



B

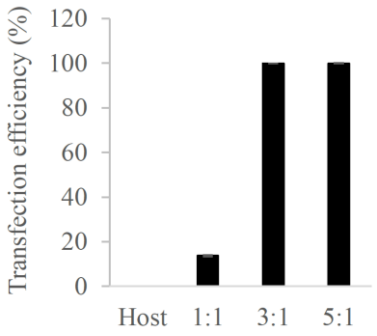


Figure S5.1. HEK 293-EBNA transfection efficiency analysis. HEK 293-EBNA cells were transfected with pCEP-PU-EmGFP plasmid (Section 9.4) using a lipofection based method (*Trans*IT-LT1; Mirus) (Section 5.2.3.). Cells transiently expressing GFP were harvested at day three post-transfection. The samples were analysed using a BD Accuri C6 Flow Cytometer with BD CSampler (BD Biosciences) using a 488 nm laser for excitation, and measuring emission 530/30 nm band pass filter and FL1 detector for green fluorescence detection. Histograms were generated by plotting fluorescence intensity versus event counts (10 000 events for each sample). (**A**) Representative histograms of untrasfected control (Host), mock transfection control (Neg Ctrl) and three different reagent:DNA transfection conditions (µL of reagent:µg of DNA; 1:1, 3:1, and 5:1). (**B**) Transfection efficiency percentage was calculated by the ratio of fluorescent cells to total number of cells analysed. Untransfected control (Host) was used to establish the autofluorescence value and the threshold ratio was set to the right edge of the control histogram (red vertical line) (Analysis in collaboration with MPhil Fu Swee Jiun student, 2014).

5.5.2 *The enhanced rHuEPO G09 variant*

We investigated the location of the four mutations in the enhanced rHuEPO G09 variant (Buchanan et al., 2012) in our full solubility screening prediction (Supplementary Table S5.1). This improved rHuEPO protein is the result of an extensive DNA library produced by applying a PCR random mutagenesis, using an error rate of eight nucleotides per thousand base pairs. This is the initial stage of the ribosome display, which next step is *In vitro* transcription of mRNA, followed by a reconstituted *E. coli* S30 extract translation system with the presence or absence of the reducing agent dithiothreitol (DTT). After translation, the ribosome-protein complexes were subjected to three selection pressures: (i) DTT as reducing agent, (ii) incubation at elevated temperature, and (iii) hydrophobic interaction chromatography matrices. The functional selection was carried out by binding to a human cognate receptor-Fc fusion partner. After four selection rounds, they found nine more active variants, where rHuEPO G09 stood out as the most active one. According to their computational analysis, using a software based on the spatial aggregation propensity (SAP) method (Accelrys Software Inc.) (Voynov et al., 2009), shown that the diminished propensity of aggregation in rHuEPO G09 variant is due to the negatively-charged substitution of glutamic acid at position 158 (G158E). This supports our hypothesis, stating that diminishing positively-charged patches by adding negatively-charged residues will generate a less prone to aggregate protein variants.

Residue			Ratio																
ALA	1	POS	1.492	LEU	35	NEG		LEU	69	NEG		ARG	103	POS	0.477	THR	137	POS	1.492
PRO	2	POS	1.492	ASN	36	HYD		LEU	70	NEG		SER	104	HYD		PHE	138	NEG	
PRO	3	HYD		GLU	37	HYD		SER	71	HYD		LEU	105	POS	0.477	ARG	139	POS	1.492
ARG	4	POS	1.492	ASN	38	HYD		GLU	72	HYD		THR	106	POS	0.477	LYS	140	POS	1.492
LEU	5	POS	1.492	ILE	39	NEG		ALA	73	NEG		THR	107	POS	0.477	LEU	141	POS	1.492
ILE	6	POS	0.003	THR	40	NEG		VAL	74	HYD		LEU	108	HYD		PHE	142	POS	1.492
CYS	7	POS	1.492	VAL	41	HYD		LEU	75	HYD		LEU	109	HYD		ARG	143	HYD	
ASP	8	HYD		PRO	42	POS	0.022	ARG	76	HYD		ARG	110	POS	0.014	VAL	144	POS	1.492
SER	9	POS	1.492	ASP	43	HYD		GLY	77	NEG		ALA	111	HYD		TYR	145	HYD	
ARG	10	HYD		THR	44	POS	1.492	GLN	78	NEG		LEU	112	NEG		SER	146	HYD	
VAL	11	HYD		LYS	45	POS	1.492	ALA	79	HYD		GLY	113	NEG		ASN	147	POS	1.492
LEU	12	HYD		VAL	46	HYD		LEU	80	NEG		ALA	114	NEG		PHE	148	POS	1.492
GLU	13	POS	1.492	ASN	47	POS	1.492	LEU	81	NEG		GLN	115	POS	0.014	LEU	149	HYD	
ARG	14	HYD		PHE	48	POS	1.492	VAL	82	NEG		LYS	116	HYD		ARG	150	POS	1.492
TYR	15	NEG		TYR	49	POS	1.492	ASN	83	HYD		GLU	117	HYD		GLY	151	POS	1.492
LEU	16	HYD		ALA	50	HYD		SER	84	HYD		ALA	118	NEG		LYS	152	POS	1.492
LEU	17	NEG		TRP	51	HYD		SER	85	NEG		ILE	119	NEG		LEU	153	HYD	
GLU	18	HYD		LYS	52	NEG		GLN	86	NEG		SER	120	NEG		LYS	154	POS	1.492
ALA	19	POS	0.477	ARG	53	HYD		PRO	87	HYD		PRO	121	NEG		LEU	155	HYD	
LYS	20	POS	1.492	MET	54	HYD		TRP	88	POS	1.492	PRO	122	NEG		TYR	156	HYD	
GLU	21	NEG		GLU	55	HYD		GLU	89	NEG		ASP	123	NEG		THR	157	POS	1.492
ALA	22	POS	0.477	VAL	56	NEG		PRO	90	HYD		ALA	124	HYD		GLY	158	POS	1.492
GLU	23	POS	1.492	GLY	57	NEG		LEU	91	HYD		ALA	125	NEG		GLU	159	POS	1.492
ASN	24	HYD		GLN	58	NEG		GLN	92	HYD		SER	126	HYD		ALA	160	NEG	
ILE	25	HYD		GLN	59	NEG		LEU	93	HYD		ALA	127	HYD		CYS	161	POS	1.492
THR	26	HYD		ALA	60	NEG		HIS	94	POS	0.477	ALA	128	HYD		ARG	162	POS	1.492
THR	27	POS	1.492	VAL	61	HYD		VAL	95	POS	0.477	PRO	129	HYD		THR	163	POS	1.492
GLY	28	HYD		GLU	62	HYD		ASP	96	NEG		LEU	130	POS	0.025	GLY	164	POS	1.492
CYS	29	POS	1.492	VAL	63	HYD		LYS	97	HYD		ARG	131	HYD		ASP	165	NEG	
ALA	30	POS	1.492	TRP	64	HYD		ALA	98	POS	0.477	THR	132	POS	0.022	ARG	166	NEG	
GLU	31	HYD		GLN	65	HYD		VAL	99	HYD		ILE	133	POS	1.492				
HIS	32	NEG		GLY	66	HYD		SER	100	HYD		THR	134	POS	1.492				
CYS	33	NEG		LEU	67	HYD		GLY	101	NEG		ALA	135	POS	1.492				
SER	34	POS	1.492	ALA	68	HYD		LEU	102	HYD		ASP	136	POS	1.492				

Table S5.1. Solubility profile of rHuEPO WT from the charged patch calculator. Complete posQ ratio output for the modified rHuEPO WT structure (PDB ID: 1EER) is shown. The largest positive patches are represented by blue (ratio > 1.0). Those proteins with ratio above 1.0 are predicted as insoluble and below 1.0 as soluble. The four stability and activity enhancing mutations in MedImmune's rHuEPO variant G09 are highlighted in red at positions 25, 27, 139 and 158. Ratio: largest positively-charged patch (posQ) value from the charged patch calculator (Chan et al., 2013). The charged mutations proposed in this thesis are highlighted in black. Charge patches: HYD, hydrophobic (non-charged); NEG, negatively-charged; POS, positively-charged.

5.5.3 *Conservation analysis of HuEPO*

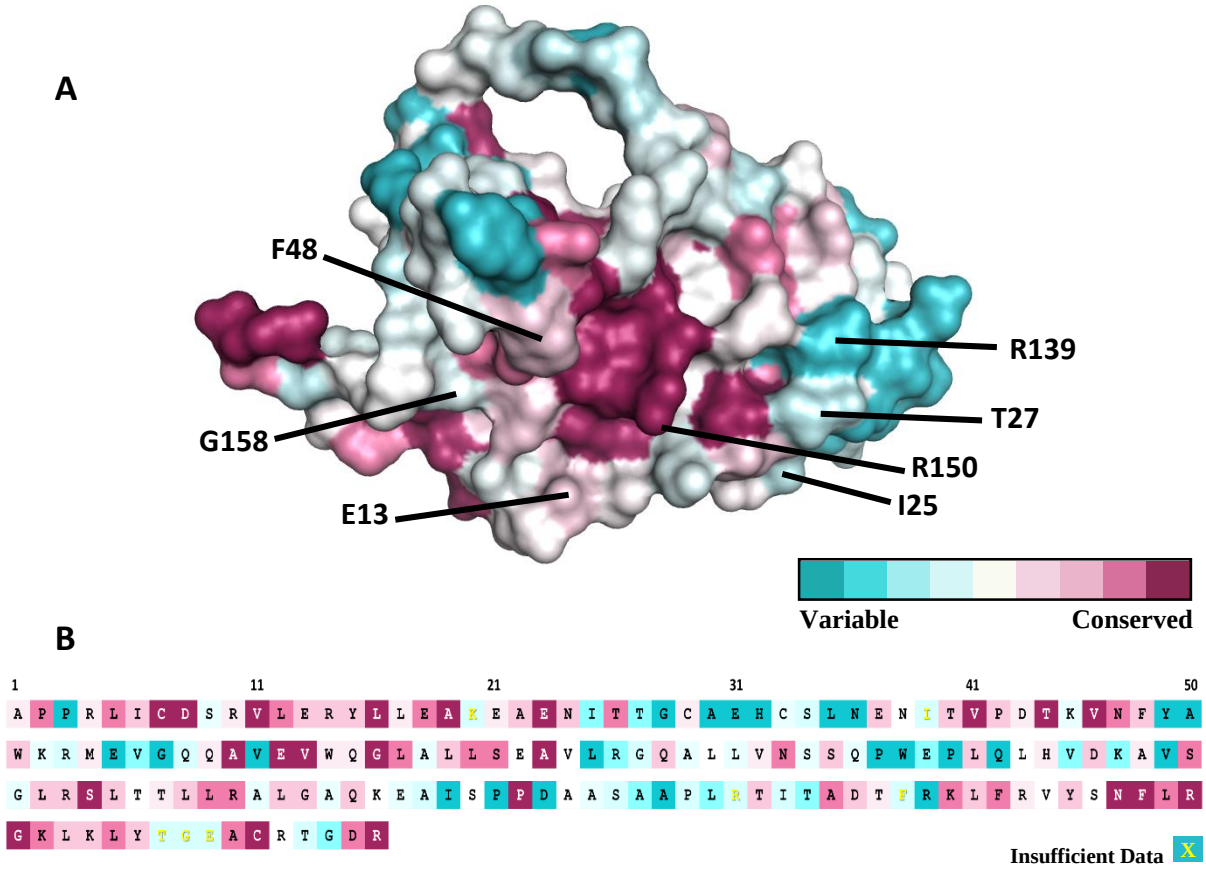


Figure S5.2. Multiple sequence alignment of rHuEPO homologues. (**A**) Sequence and surface map coloured by residue conservation scores of rHuEPO wild-type. Target residues in al rHuEPO variants are highlighted. The structure was rendered using PyMOL. (**B**) Panel showing the sequence conservation of HuEPO WT. Conservation was calculated using ConSurf server with default parameters (Ashkenazy et al., 2010). This default parameters consist suing the UniRef90 database (Suzek et al., 2007), which removes redundancy at 90% sequence identity. In addition, the default number of mohomologues sequences was set to 150 with a Position-Specific Iterated BLAST (PSI-BLAST) E-value to 0.0001 to minimise the chance of including non-homologues.

5.5.4 *Glycosylation and charge patches distribution on rHuEPO WT surface*

We have structurally explored the location of the three *N*-glycosylated sites of HuEPO WT in relation to the calculated charged surfaces patches from our algorithm (Supplementary Figure S5.3). We have found that none of the mutations (i.e. sties 13, 48 and 150) interfere with the glycosylation areas of rHuEPO. Two of the *N*-glycosylation sites (Asn-38 and Asn-83) are located above the interaction sites with both receptors and one (Asn-24) just below in between the receptors (Supplementary Figure S5.3). Most of these coverages are around the positively-charged (blue) on receptor 1 (R131-V144) and receptor 2 (S71, L75, R103, T106 and T107) and a small negative and positive charge patch (L17, E21, I25 and H94) below the interactions.

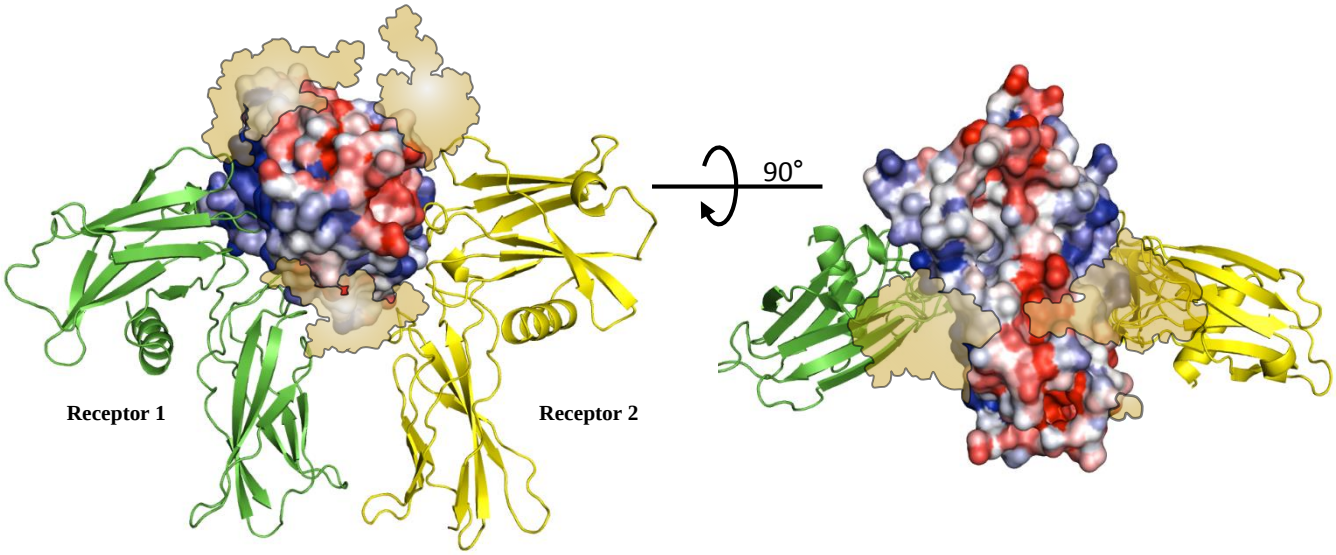


Figure S5.3 Glycosylation and charge patches distribution on rHuEPO WT surface with HuEPO receptors (PDB ID: 1EER). The electrostatic potential of rHuEPO WT surface was computed in the algorithm developed in our group (Chan et al., 2013). The three *N*-glycosylation sites are represented from a tetraantennary sialoglycan unit in a tetra-antennary carbohydrate model based on Elliot et al. (Elliott et al., 2003, Elliott, 2009). Blue and red coloured areas represent positive and negative charges, respectively.

5.5.5 *Is the carboxyl-terminal His-Tag removed from mature rHuEPO?*

Since we are investigating the consequences of amino acid substitutions, mainly positively and negatively-charged residues, we decided to leave the last R166 amino acid intact at the C-terminal preceding the 6xHis-tag. The removal of this residue in the mature form of rHuEPO from mammalian cells has been studied previously (Recny et al., 1987). Debeljak and co-workers deleted the R166 residue, hypothesising a removal of the subsequent 6x his-tag at the c-terminal in the mature form (Debeljak et al., 2006). As they reported, we could not detect rHuEPO by using a mouse anti-polyHis (Sigma), which we have previously reported to detect *Escherichia coli*-derived non-glycosylated rHuEPO (Carballo-Amador et al., 2014b). In order to investigate the presence of the His-Tag, we applied a nickel immobilized metal chelate affinity chromatography.

5.5.5.1 *Purification of C-terminal His-tagged rHuEPO WT and variants*

Secreted rHuEPO from HEK 293-EBNA cell cultures was subjected to purification using His60 Ni Superflow^{TM} resin (Clontech). Medium samples (500 µl) from day 7 containing the his-tagged rHuEPOs was added to individual gravity columns (previously treated with equilibration buffer, 50 mM sodium phosphate, 300 mM sodium chloride, 20 mM imidazole, pH 7.4). Samples were collected from the flow-through and from the wash step (wash buffer, 50 mM sodium phosphate, 300 mM sodium chloride, 40 mM imidazole, pH 7.4). rHuEPO purification was achieved by addition of elution buffer (50 mM sodium phosphate, 300 mM sodium chloride, 300 mM imidazole, pH 7.4). rHuEPO purification was analysed by SDS-PAGE protein staining and Western blot (Section 5.2.6). For protein staining, gels were rinsed with Milli-Q water to remove any excess electrode running buffer and were immersed in

InstantBlue (Expedeon) and incubated for 30 min at room temperature with gently shaking. Then, the stain was removed and the gel was rinsed with Milli-Q water and subjected to analysis with the Odyssey Imaging System (LI-COR). Detection of partial purified rHuEPO WT and variants was possible by blotting with a mouse anti-HuEPO monoclonal antibody (R&D Systems; MAB287) (1:1500 dilution). As shown in figure S5.2, no signal was detected in the unbound and during washing fractions, only in the elution step by western blot and protein staining, indicating that the C-terminal his-tag is not removed from mature rHuEPO.

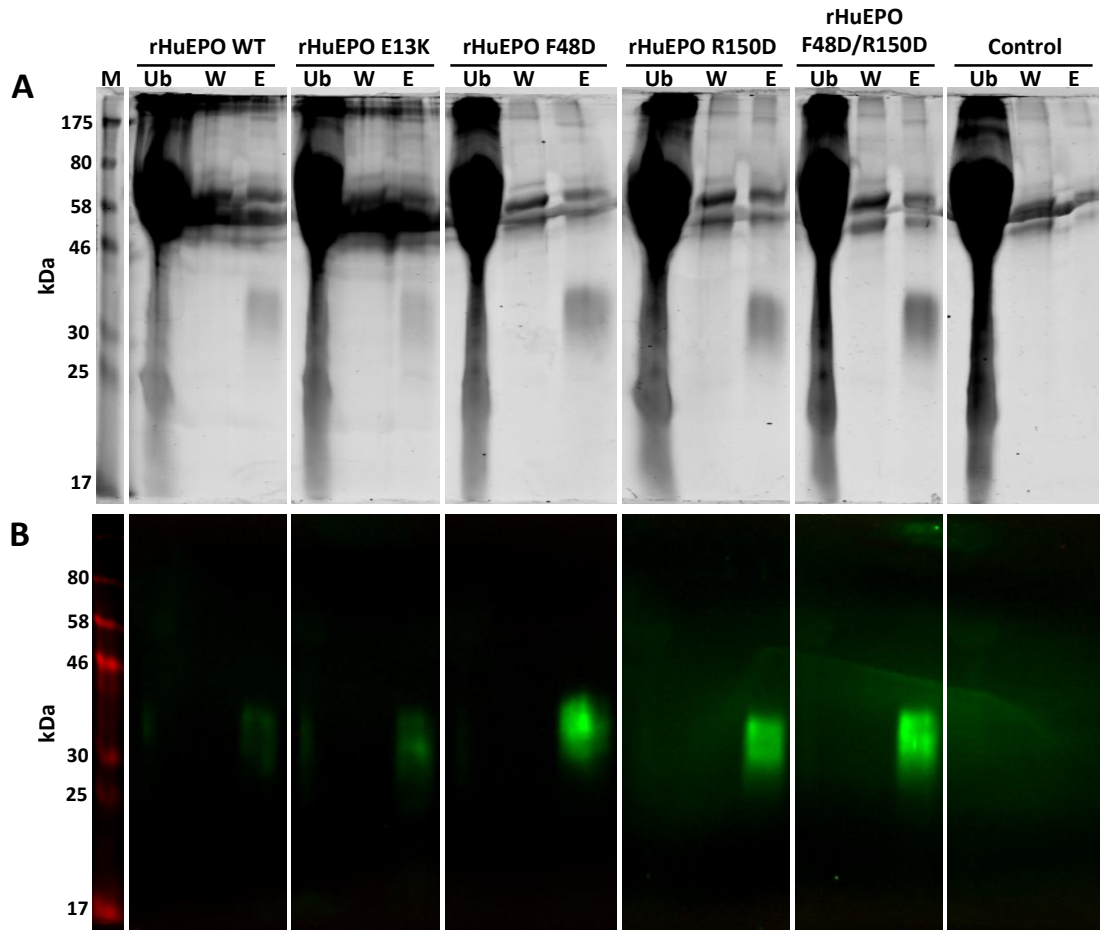


Figure S5.4. Analysis of the purification of rHuEPO WT and variants. (**A**) SDS-PAGE protein staining and (**B**) western blot analysis of purification steps. Lane Ub, flow-through fraction collected after applying media samples on His60 Ni Superflow^{TM} resin. Lane W, fraction obtained after applying washing buffer (50 mM sodium phosphate, 300 mM sodium chloride, 40 mM imidazole, pH 7.4). Lane E, rHuEPO eluted from His60 Ni Superflow^{TM} resin by addition of elution buffer (50 mM sodium phosphate, 300 mM sodium chloride, 300 mM imidazole, pH 7.4). M, prestained SDS-PAGE marker (Bio-Rad); Ub, Unbind; W, Wash; E, Elution.

Acknowledgements

We would like to thank Dr. E. McKenzie for supply host cells and expression vector, Prof. S. High and MPhil J. Fu for discussion, and also CONACyT for contributing PhD funds.

5.6 References

ASHKENAZY, H., EREZ, E., MARTZ, E., PUPKO, T. & BEN-TAL, N. 2010. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res,* 38**,** W529-33.

BRINKS, V., HAWE, A., BASMELEH, A. H., JOACHIN-RODRIGUEZ, L., HASELBERG, R., SOMSEN, G. W., JISKOOT, W. & SCHELLEKENS, H. 2011. Quality of original and biosimilar epoetin products. *Pharm Res,* 28**,** 386-93.

BUCHANAN, A., FERRARO, F., RUST, S., SRIDHARAN, S., FRANKS, R., DEAN, G., MCCOURT, M., JERMUTUS, L. & MINTER, R. 2012. Improved drug-like properties of therapeutic proteins by directed evolution. *Protein Eng Des Sel,* 25**,** 631-8.

CARBALLO-AMADOR, M. A., WARWICKER, J. & DICKSON, A. J. 2014. Increasing solubility in recombinant erythropoietin through modification of surface patches. *in preparation*.

CHAN, P., CURTIS, R. A. & WARWICKER, J. 2013. Soluble expression of proteins correlates with a lack of positively-charged surface. *Sci Rep,* 3**,** 3333.

CLOSE, D. W., DON PAUL, C., LANGAN, P. S., WILCE, M. C. J., TRAORE, D. A. K., HALFMANN, R., ROCHA, R. C., WALDO, G. S., PAYNE, R. J., RUCKER, J. B., PRESCOTT, M. & BRADBURY, A. R. M. 2014. TGP, an extremely stable, non-aggregating fluorescent protein created by structure-guided surface engineering. *Proteins: Structure, Function, and Bioinformatics***,** n/a-n/a.

DAVIS, J. M., ARAKAWA, T., STRICKLAND, T. W. & YPHANTIS, D. A. 1987. Characterization of recombinant human erythropoietin produced in Chinese hamster ovary cells. *Biochemistry,* 26**,** 2633-8.

DEBELJAK, N., FELDMAN, L., DAVIS, K. L., KOMEL, R. & SYTKOWSKI, A. J. 2006. Variability in the immunodetection of His-tagged recombinant proteins. *Anal Biochem,* 359**,** 216-23.

DEREWENDA, Z. 2010. Application of protein engineering to enhance crystallizability and improve crystal properties. *Acta Crystallographica Section D,* 66**,** 604-615.

ELLIOTT, S. 2009. New molecules and formulations. *In:* ELLIOTT, S., FOOTE, M. & MOLINEUX, G. (eds.) *Erythropoietins, Erythropoietic Factors, and Erythropoiesis.* Birkhäuser Basel.

ELLIOTT, S., LORENZINI, T., ASHER, S., AOKI, K., BRANKOW, D., BUCK, L., BUSSE, L., CHANG, D., FULLER, J., GRANT, J., HERNDAY, N., HOKUM, M., HU, S., KNUDTEN, A., LEVIN, N., KOMOROWSKI, R., MARTIN, F., NAVARRO, R., OSSLUND, T., ROGERS, G., ROGERS, N., TRAIL, G. & EGRIE, J. 2003. Enhancement of therapeutic protein in vivo activities through glycoengineering. *Nat Biotechnol,* 21**,** 414-21.

ELLIOTT, S., LORENZINI, T., CHANG, D., BARZILAY, J. & DELORME, E. 1997. Mapping of the active site of recombinant human erythropoietin. *Blood,* 89**,** 493-502.

FOTIOU, F., ARAVIND, S., WANG, P. P. & NERAPUSEE, O. 2009. Impact of illegal trade on the quality of epoetin alfa in Thailand. *Clin Ther,* 31**,** 336-46.

FRADET-TURCOTTE, A., CANNY, M. D., ESCRIBANO-DIAZ, C., ORTHWEIN, A., LEUNG, C. C., HUANG, H., LANDRY, M. C., KITEVSKI-LEBLANC, J., NOORDERMEER, S. M., SICHERI, F. & DUROCHER, D. 2013. 53BP1 is a reader of the DNA-damage-induced H2A Lys 15 ubiquitin mark. *Nature,* 499**,** 50-4.

GASTEIGER, E., HOOGLAND, C., GATTIKER, A., DUVAUD, S. E., WILKINS, M., APPEL, R. & BAIROCH, A. 2005. Protein Identification and Analysis Tools on the ExPASy Server. *In:* WALKER, J. (ed.) *The Proteomics Protocols Handbook.* Humana Press.

GATTI, M., PINATO, S., MASPERO, E., SOFFIENTINI, P., POLO, S. & PENENGO, L. 2012. A novel ubiquitin mark at the N-terminal tail of histone H2As targeted by RNF168 ubiquitin ligase. *Cell Cycle,* 11**,** 2538-44.

HIGUCHI, M., OH-EDA, M., KUBONIWA, H., TOMONOH, K., SHIMONAKA, Y. & OCHI, N. 1992. Role of sugar chains in the expression of the biological activity of human erythropoietin. *J Biol Chem,* 267**,** 7703-9.

JÄCKEL, C., KAST, P. & HILVERT, D. 2008. Protein Design by Directed Evolution. *Annual Review of Biophysics,* 37**,** 153-173.

JACOBS, K., SHOEMAKER, C., RUDERSDORF, R., NEILL, S. D., KAUFMAN, R. J., MUFSON, A., SEEHRA, J., JONES, S. S., HEWICK, R., FRITSCH, E. F. & ET AL. 1985. Isolation and characterization of genomic and cDNA clones of human erythropoietin. *Nature,* 313**,** 806-10.

JELKMANN, W. 2013. Physiology and pharmacology of erythropoietin. *Transfus Med Hemother,* 40**,** 302-9.

JEONG, Y. T., CHOI, O., LIM, H. R., SON, Y. D., KIM, H. J. & KIM, J. H. 2008. Enhanced sialylation of recombinant erythropoietin in CHO cells by human glycosyltransferase expression. *J Microbiol Biotechnol,* 18**,** 1945-52.

KAMIONER, D. 2012. Erythropoietin biosimilars currently available in hematology-oncology. *Target Oncol,* 7 Suppl 1**,** S25-8.

LAI, P. H., EVERETT, R., WANG, F. F., ARAKAWA, T. & GOLDWASSER, E. 1986. Structural characterization of human erythropoietin. *J Biol Chem,* 261**,** 3116-21.

LISOWSKA, E. 2002. The role of glycosylation in protein antigenic properties. *Cell Mol Life Sci,* 59**,** 445-55.

MATTIROLI, F. & SIXMA, T. K. 2014. Lysine-targeting specificity in ubiquitin and ubiquitin-like modification pathways. *Nat Struct Mol Biol,* 21**,** 308-16.

MATTIROLI, F., UCKELMANN, M., SAHTOE, D. D., VAN DIJK, W. J. & SIXMA, T. K. 2014. The nucleosome acidic patch plays a critical role in RNF168-dependent ubiquitination of histone H2A. *Nat Commun,* 5.

MCKOY, J. M., STONECASH, R. E., COURNOYER, D., ROSSERT, J., NISSENSON, A. R., RAISCH, D. W., CASADEVALL, N. & BENNETT, C. L. 2008. Epoetin-associated pure red cell aplasia: past, present, and future considerations. *Transfusion,* 48**,** 1754-62.

MIKHAIL, A. & FAROUK, M. 2013. Epoetin biosimilars in Europe: five years on. *Adv Ther,* 30**,** 28-40.

NARHI, L. O., ARAKAWA, T., AOKI, K. H., ELMORE, R., ROHDE, M. F., BOONE, T. & STRICKLAND, T. W. 1991. The effect of carbohydrate on the structure and stability of erythropoietin. *J Biol Chem,* 266**,** 23022-6.

NIWA, T., YING, B. W., SAITO, K., JIN, W., TAKADA, S., UEDA, T. & TAGUCHI, H. 2009. Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of Escherichia coli proteins. *Proc Natl Acad Sci U S A,* 106**,** 4201-6.

PARK, S. S., PARK, J., KO, J., CHEN, L., MERIAGE, D., CROUSE-ZEINEDDINI, J., WONG, W. & KERWIN, B. A. 2009. Biochemical assessment of erythropoietin products from Asia versus US Epoetin alfa manufactured by Amgen. *Journal of Pharmaceutical Sciences,* 98**,** 1688-1699.

PAROUTIS, P., TOURET, N. & GRINSTEIN, S. 2004. The pH of the secretory pathway: measurement, determinants, and regulation. *Physiology (Bethesda),* 19**,** 207-15.

PRADITPORNSILPA, K., TIRANATHANAGUL, K., KUPATAWINTU, P., JOOTAR, S., INTRAGUMTORNCHAI, T., TUNGSANGA, K., TEERAPORNLERTRATT, T., LUMLERTKUL, D., TOWNAMCHAI, N., SUSANTITAPHONG, P., KATAVETIN, P., KANJANABUCH, T., AVIHINGSANON, Y. & EIAM-ONG, S. 2011. Biosimilar recombinant human erythropoietin induces the production of neutralizing antibodies. *Kidney Int,* 80**,** 88-92.

RATANJI, K. D., DERRICK, J. P., DEARMAN, R. J. & KIMBER, I. 2014. Immunogenicity of therapeutic proteins: influence of aggregation. *J Immunotoxicol,* 11**,** 99-109.

RECNY, M. A., SCOBLE, H. A. & KIM, Y. 1987. Structural characterization of natural human urinary and recombinant DNA-derived erythropoietin. Identification of des-arginine 166 erythropoietin. *J Biol Chem,* 262**,** 17156-63.

SCHRÖDINGER, L. L. C. 2010. The PyMOL Molecular Graphics System, Version 1.3r1.

SEIDL, A., HAINZL, O., RICHTER, M., FISCHER, R., BOHM, S., DEUTEL, B., HARTINGER, M., WINDISCH, J., CASADEVALL, N., LONDON, G. M. & MACDOUGALL, I. 2012. Tungsten-induced denaturation and aggregation of epoetin alfa during primary packaging as a cause of immunogenicity. *Pharm Res,* 29**,** 1454-67.

SKIBELI, V., NISSEN-LIE, G. & TORJESEN, P. 2001. *Sugar profiling proves that human serum erythropoietin differs from recombinant human erythropoietin*.

SU, D., ZHAO, H. & XIA, H. 2010. Glycosylation-modified erythropoietin with improved half-life and biological activity. *Int J Hematol,* 91**,** 238-44.

SUZEK, B. E., HUANG, H., MCGARVEY, P., MAZUMDER, R. & WU, C. H. 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics,* 23**,** 1282-1288.

THOMAS, P. & SMART, T. G. 2005. HEK293 cell line: A vehicle for the expression of recombinant proteins. *Journal of Pharmacological and Toxicological Methods,* 51**,** 187-200.

VOYNOV, V., CHENNAMSETTY, N., KAYSER, V., HELK, B. & TROUT, B. L. 2009. Predictive tools for stabilization of therapeutic proteins. *mAbs,* 1**,** 580-582.

WANG, W., SINGH, S. K., LI, N., TOLER, M. R., KING, K. R. & NEMA, S. 2012. Immunogenicity of protein aggregates--concerns and realities. *Int J Pharm,* 431**,** 1-11.

WEN, D., BOISSEL, J. P., SHOWERS, M., RUCH, B. C. & BUNN, H. F. 1994. Erythropoietin structure-function relationships. Identification of functionally important domains. *J Biol Chem,* 269**,** 22839-46.

WETZEL, R. 1996. For Protein Misassembly, It's the "I" Decade. *Cell,* 86**,** 699-702.

Chapter 6

Paper 4:

Alteration of lysine and arginine content as a strategy to modify protein solubility: a test for *E. coli* proteins

Carballo-Amador M.A.[12], Dickson A.J.[1] and Warwicker J.[2]

[1] Faculty of Life Sciences, University of Manchester, Michael Smith Building, Oxford Road, Manchester M13 9PT, UK

[2] Faculty of Life Sciences, University of Manchester, Manchester Institute of Biotechnology, 131 Princess Street, Manchester M1 7DN, UK

Abstract

Protein aggregation is an undesired physicochemical mechanism whether for biophysical and structural studies or for biopharmaceutical companies, at any scale. In *Escherichia coli*, protein accumulation in the cytoplasm can result in protein aggregation to form what are known as inclusion bodies (IBs). Several experimental approaches have been undertaken to prevent protein aggregation. However, there is no universal approach or technology that solves protein aggregation. Recently, our group found that the sequence-based property of lysine *versus* arginine content separated *E. coli* proteins by solubility. In this study, we investigated solubility alterations for three highly soluble *E. coli* proteins (thioredoxin-1 [TRX], cold shock-like protein cspB [cspB], and the histidine-containing phosphocarrier protein [HPr]), with varying degree of lysine substitution by arginine. These experiments are predicted to decrease the solubility of the variants, according to our computational calculations. Our findings revealed a significant decrease in solubility for cspB and HPr, which is more evident in variants with low or null lysine content. However, for the expression of TRX variants, solubility only falls under low induction conditions (low temperature and IPTG inducer) compared to WT. This computational and experimental approach is a first step in studying to what extent the lysine:arginine ratio modifies solubility.

6.1 Introduction

Proteins have a wide range of properties as a result of natural evolution from simple cells to complex organisms. These properties are encrypted in their own primary sequence information (Berg et al., 2002). Once the polypeptide is translated the next step is to achieve its native structure form during the folding process. It is here where one of the main challenges for the recombinant protein production at any scale makes its appearance, protein aggregation. This phenomenon has been described as the consequence of partial or unfolded protein interactions (Mitraki and King, 1989, Fink, 1998). In *Escherichia coli*, protein accumulation in the cytoplasm can result in protein aggregation to form what are known as inclusion bodies (Kane and Hartley, 1988). This process often occurs during overexpression of eukaryotic-sourced proteins, proteins which should carry post-translational modifications that do not occur in the cytosol of *E. coli* (Marston, 1986, Fischer et al., 1993, Sahdev et al., 2008). Some of these proteins are market-leading biotherapeutics, hence a basic molecular understanding is essential in order to achieved production of cost-efficient biotherapeutics.

Several experimental approaches have been undertaken to prevent protein aggregation in *E. coli*. These have included lowering of cultivation temperature (Schein and Noteborn, 1988) and inducer concentration (Weickert et al., 1996), co-expression of molecular chaperones (Mogk et al., 2002), use of soluble fusion-tags (Davis et al., 1999), and modification of codon usage (Gustafsson et al., 2004, Hatfield and Roth, 2007). In addition, during the last two decades computational approaches have been developed to aid prediction of aggregation propensity based on the amino acid sequence analysis, e.g. AGGRESCAN (Conchillo-Sole et al., 2007), PASTA 2.0 (Walsh et al., 2014), TANGO (Fernandez-Escamilla et al., 2004), and Zyggregator (Tartaglia and Vendruscolo, 2008). An advantage for each of these prediction packages is that the sequence based input is readily available but this

advantage is countered by a lack of consideration of three-dimensional structural information. For example, non-polar regions that can lead to association are buried inside folded protein structure, hence attempts to mutate these regions to prevent aggregation could result in loss of the native fold. Therefore, approaches are also needed that consider fold and structure, where available. An extensive experimental protein solubility study on cell-free expression of *E. coli* proteins (Niwa et al., 2009) has opened the field for further work on protein solubility issues (Agostini et al., 2012, Samak et al., 2012, Chan et al., 2013, Fang and Fang, 2013, Agostini et al., 2014, Klus et al., 2014, Warwicker et al., 2014). Our group has used the *E. coli* protein solubility data to discern both structure- and sequence-based features that best separate the most and least soluble protein subsets (Chan et al., 2013, Warwicker et al., 2014). The stand-out discriminatory feature was the size of the largest positively-charged patch on the protein surface (Chan et al., 2013), which has been explored experimentally (Carballo-Amador et al., 2014a, Carballo-Amador et al., 2014b). It is important to note that our working hypothesis for the mechanism by which positive charge patch size could mediate solubility is specific to expression systems, which are rich in negatively-charged macroions (mRNAs). The current Chapter considers a property that we feel may be more general, relating to protein solubility in different environments. When we looked at the make-up of positive charges in proteins (largely lysine and arginine side chain charges at physiological pH), it was found that the sequence-based property of lysine (K) *versus* arginine (R) content also discriminatory between the least and more soluble subsets of *E. coli* protein (Warwicker et al., 2014). A greater lysine:arginine content was associated with the more soluble subset. Furthermore, we found proteins at high naturally occurring concentrations (*versus* those at lower concentrations) also had a relative preference for lysine (Warwicker et al., 2014), which lead to the suggestion that this balance of lysine and arginine could be a general effect, not just specific to expression systems. In

support of this, an improvement in protein stability has been reported by changing lysines to arginines in green fluorescent protein (GFP) (Sokalingam et al., 2012).

The current investigation focuses on the exploration of *E. coli* K-12 proteome, testing the role of K to R ratio in protein expression in the first instance. At this stage, testing has not progressed to biophysical characterisation of purified protein. To select proteins, 2,931 *E. coli* proteins were sorted by highest calculated K to R ratio proteins, and cross-referenced with proteins observed experimentally to be highly soluble in cell-free expression (Niwa et al., 2009). Three proteins were selected, thioredoxin-1 (TRX), cold shock-like protein cspB (cspB), and the histidine-containing phosphocarrier protein (HPr). TRX is a small (108 amino acid) redox protein involved in varied cellular functions, including ribonucleotide reduction (Thelander, 1967) and protein folding (Yasukawa et al., 1995). HPr is a small peptide (85 amino acid) component of the phosphoenolpyruvate (PEP): carbohydrate phosphotransferase system (PTS) (Meadow et al., 1990). CspB (71 residues) is one of the major cold-shock proteins that is triggered by a temperature downshift. These proteins are essential for *E.coli* cell growth under sub-optimal temperatures (Etchegaray and Inouye, 1999).

In order to investigate the consequences of changes in K to R ratio for solubility when these three proteins was expressed in *E. coli*, we designed intermediate (C2) and partial or full substitution constructs (C3) (i.e. low or null lysine content). We have observed a significant drop in solubility for cspB and HPr compared to wild-type versions, which is more prominent in the low or null lysine content construct. This computational and experimental approach shall contribute to rationalisation of means to improve yields in recombinant protein technologies, especially in relation to design of novel format biotherapeutics.

6.2 Materials and methods

6.2.1 *Lysine-Arginine ratio screening in E. coli proteins*

Mutating lysines to arginines represents an experimental continuation of some initial computational observations published recently from our group (Warwicker et al., 2014). A sequence-based analysis was performed for K to R ratio ranking from a solubility database (Niwa et al., 2009) of 2,931 *E. coli* proteins. The selection filter of the proteins to be mutated were designated by the top soluble proteins according to Taguchi's group classification (Niwa et al., 2009), high K to R ratio, characterised proteins, relatively small protein size (< 200 aa) for easier gene construction and expression, predicted folded, without transmembrane domains, and to form a non-redundant set (i.e. no homologues). Based on these parameters three candidate proteins were selected: the small electron-transfer thioredoxin-1 protein (TRX), cold shock-like protein cspB, and histidine-containing phosphocarrier protein (HPr). Two variants for each protein were evaluated, in addition to the unmutated protein, an intermediate and a partial or full substitution construct (i.e. low or null lysine content).

6.2.2 *Computational structural analysis*

In order to avoid removing structurally or functionally important lysines the 3D structures were analysed for TRX (PDB ID: 2TRX), HPr (PDB ID: 2JEL) and an 84% cspB homologue (PDB ID: 2L15). All structural analysis were carried out by using the open sources Swiss-PdbViewer 4.0.1 (Guex and Peitsch, 1997) and PyMOL Molecular Graphics System version 1.3 (Schrödinger, 2010). All the positively-charged polar residues (K and R) are on the surface of the proteins. Patch calculations were carried out using an algorithm developed in our group

(Chan et al., 2013). Multiple sequence alignments were performed using the ConSurf server (Ashkenazy et al., 2010, Celniker et al., 2013) and visualised using Jalview, a multiple alignment editor (Waterhouse et al., 2009). The Fold Index tool was used to predict whether a protein folds (Prilusky et al., 2005). Trans-membrane domains examination was carried out by analysis of annotation in the UniProt database (Consortium, 2014).

6.2.3 *Construction of expression vectors*

Each gene encoding a target protein (wild-type, C2 and C3) was synthesised by GeneArt (Life Technologies). Alterations in codon usage optimisation were not carried out, since the expression is within an endogenous system. Enzymatic restriction sites flanking the genes for *BamHI* and *EcoRI* were included to facilitate the insertion into pHis vector (Section 9.2), which was kindly provided by Dr. Edward McKenzie of the University of Manchester. This expression plasmid is a modified version of the commercial pET-16b vector (Novagen) under the T7 promoter. The gene sequence for each plasmid was as follows: 5'-6xHis-Thrombin cleavage site-*BamHI*-Target Gene-*EcoRI*-3'.

6.2.4 *Protein expression and solubility assay*

All proteins were expressed in *E. coli* BL21 (DE3) pLysS strain. Transformed cells with the pHis-Target-Gene plasmids were grown overnight in 5 ml working volume of Luria-Bertani (LB) medium (10g tryptone, 5g yeast extract, 5g NaCl) containing 100µg/ml ampicillin and 50µg/ml chloramphenicol at 37ºC with shaking at 220 rpm. These antibiotics were added in order to preserve the pHis and pLysS plasmid, respectively. Next day, 1 ml of pre-culture was

transferred to 50 ml (2% [v/v]) LB supplemented with 2% (w/v) glucose with 100µg/ml ampicillin in 250 ml shake flasks in triplicate biological replicates. Cells were grown to an OD_{600} of 0.6 to 0.8 at constant temperature of 37°C with shaking at 180 rpm. Protein expression was induced with 0.5 mM IPTG. In addition, TRX and HPr expression at low induction conditions was carried out with a constant temperature of 25°C and induction with 0.05 mM IPTG. After growing for 5 h post-induction, cells were harvested by centrifugation at 6,500 g for 15 min at 4°C. Bacterial pellets were suspended in 5 ml of lysis buffer (25 mM Tris pH 7.5, 150 mM NaCl, 1% Triton X-100) and were stored at -20ºC for further use. The cell pellets were subjected to sonication for cell disruption by six cycles of 30 s at 20% amplitude and then allowed to cool for 30 s on ice water bath. Separation of soluble and total fractions was performed by centrifugation at 18,000 g for 30 min at 4ºC of 1 ml of each sample from the whole cell lysate. The supernatants were collected and handled as the soluble fraction. Uncentrifuged samples were handled as total fraction. Protein solubility (%) was calculated as the densitometric ratio of soluble to total fraction.

6.2.5 *SDS-PAGE and Western blot*

Separation of soluble and total fractions was analysed by sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE) in a 12% (w/v) separating gel with a 5% (w/v) stacking gel using the Mini-PROTEAN Tetra Cell (BioRad). Samples containing equal volumes of protein extracts were subjected to heating at 95°C by 5 min in 6x denaturing buffer (375 mM Tris pH 6.8, 12% [w/v] SDS, 60% [v/v] glycerol, 0.06% [w/v] bromophenol blue, 5.5% [v/v] β-mercaptoethanol). Proteins were separated in electrode running buffer (50 mM Tris, 0.38 M glycine, 0.2% [w/v] SDS) at 60 V until samples migrated into the separating gel and then the voltage was increased to 160 V at room temperature. For the specific detection of

the recombinant proteins, proteins were transferred to nitrocellulose membrane surrounded by transfer pads (BIO-RAD) that were soaked into blotting buffer (25 mM Tris, pH7.4, 0.2 M glycine and 20% [v/v] methanol) after their separation by SDS-PAGE. The transfer was performed using a transblot semi-dry transfer cell (Bio-Rad) at 15V for 45 min. After blocking the membrane in blocking buffer (5% [w/v] skimmed milk in TBS-Tween pH 7.4) for 12-14 h at 4°C with gentle shaking, the membrane was incubated for 2 h in blocking buffer solution containing mouse anti-polyHis (Sigma) (1:5000 dilution) at room temperature in rotation. The primary antibody was removed and the membrane was washed three times (5 min each time) at room temperature in TBS-Tween. For detection of the protein bands, an IR-labeled secondary Donkey anti-Mouse IgG antibody (LI-COR) (1:15000 dilution) in blocking buffer solution was added for 45 min. Followed the incubation, the secondary antibody was removed and the membrane was washed three times as mentioned previously. For IR detection, blots were imaged with the Odyssey Imaging System. Bands were quantified in the Image Studio Lite software (LI-COR) in order to estimate the protein solubility and relative total expression.

6.3 Results

6.3.1 *Selection of proteins and design of lysine to arginine mutations*

A recently published analysis suggested that there was a correlation between the relative content of lysines and arginines and protein solubility (Warwicker et al., 2014), with lysines more favourable for protein solubility than arginines. This hypothesis was used to identify candidate proteins for solubility analysis in expression in *E. coli*. A summary of the sequence-

based analysis of K to R ratio of 2,931 *E. coli* proteins is shown in Table 6.1. From this analysis, the selected proteins were TRX, HPr and cspB. These proteins scored within our selection parameters, ranging from 71 to 108 residues, lack of trans-membrane domains, high soluble percentage (Niwa et al., 2009), and are predicted to be favourable for folding. The selected proteins were targeted for lysine to arginine substitutions, in order to examine the consequences for solubility. For each protein a sequence analysis of wild-type and two constructs was performed in order to highlight the variability (Fig. 6.1A). In addition, the multiple sequence alignment of 150 homologues for each protein showed a high degree of variation for most K and R residues (Fig. 6.1B and C). Following previous work from our group (Chan et al 2013), and experimental analysis (Carballo-Amador et al., 2014a, Carballo-Amador et al., 2014b), the largest positively-charged patches for each of the three proteins were computed, and the ratio to a solubility (in expression) threshold evaluated. These ratios were 0.22 (TRX), 0.63 (HPr), and 0.52 (cspB), i.e. all below 1.0 and predicted to be soluble. Sizeable changes in charge patches, or non-polar patches, would not be expected for the lysine to arginine mutants. This consistency is borne out in Fig. 6.2 (charge patches) and Fig. 6.3 (non-polar patches).

TABLE 6.1. Lysine-Arginine ratio screening in *E. coli* proteins.

Protein	Gene	K/R	Niwa solub	Residues (aa)	TM domains	Fold Index
Uncharacterised	ygaU	17	90%	149	No	0.086
Uncharacterised	yhfU	13	90%	117	No	0.263
50S ribosomal protein	rplL	13	99%	121	No	0.267
Protein PhnA	yjdM	13	102%	111	No	0.032
Protein RacC	racC	11	105%	91	No	0.185
Thiol:disulfide interchange proteins	dsbC	11	82%	236	No	0.160
Sec-independent protein translocase	tatE	10	127%	67	Yes	0.250
Thioredoxin-1	**trxA**	**10**	**89%**	**108**	**No**	**0.197**
N-acetylneuraminate epimerase	nanM	7.5	76%	368	No	0.108
Uncharacterized fimbrial-like protein	ygiL	7	92%	183	No	0.082
Phosphocarrier protein HPr	**ptsH**	**7**	**88%**	**85**	**No**	**0.167**
Uncharacterized	yehD	7	96%	180	No	0.163
Uncharacterized	yidB	7	75%	132	No	0.187
Uncharacterized	yadN	6.5	77%	194	No	0.145
Cold shock-like protein	**cspB**	**6**	**114%**	**71**	**No**	**0.214**

Summary of the proteins with highest K/R ratio among 2, 931 *E. coli* proteins. TM (transmembrane) domains were investigated using the UniProt database (Consortium, 2014). Positive values via Fold Index predictor suggest a folded state of the target protein.

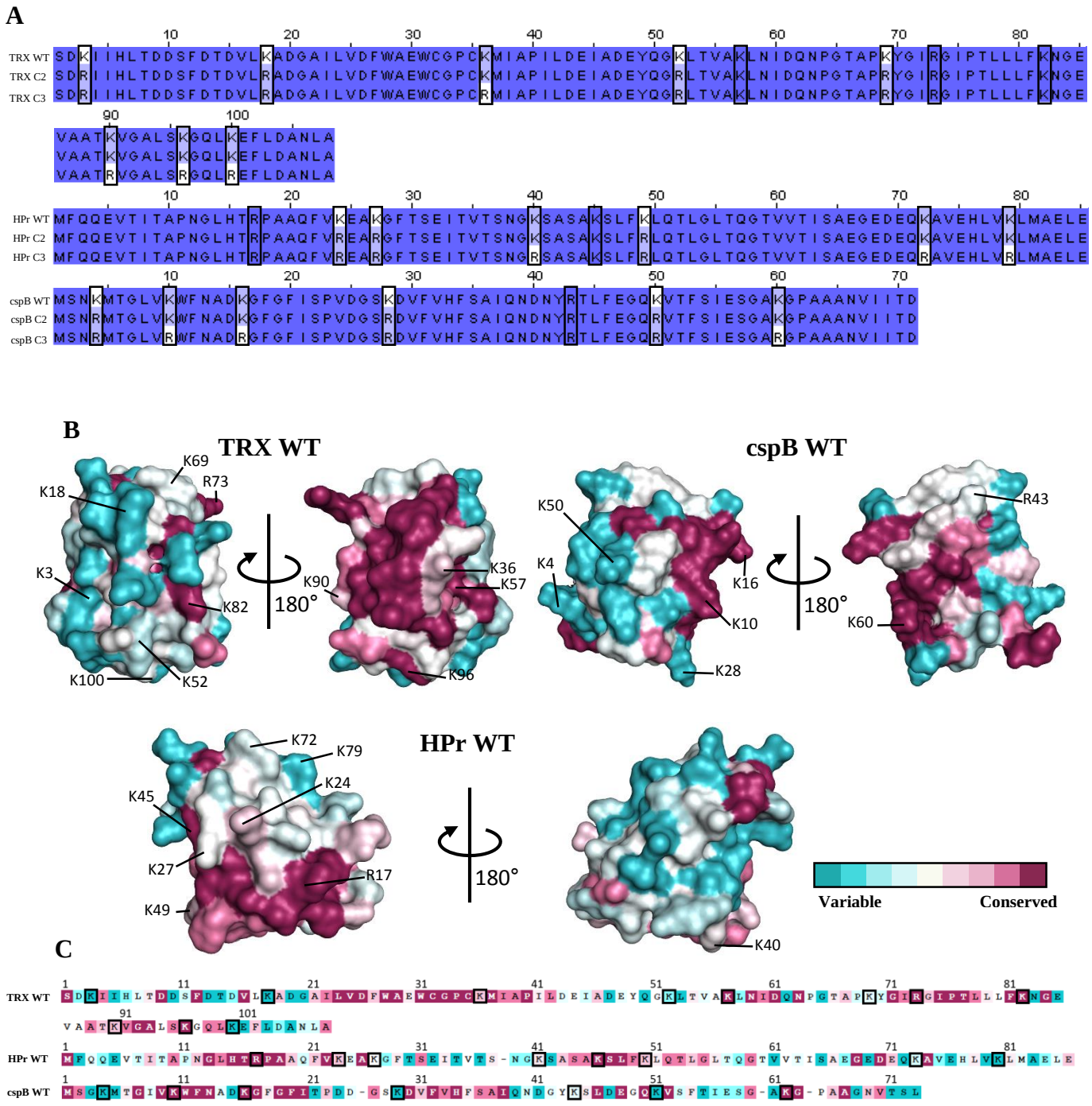


Fig. 6.1. Sequence alignment and conservation analysis of the three selected *E. coli* proteins. (**A**) Amino acids were coloured by percentage identity in Jalview. (**B**) K-R residues location on the protein surface, demosntrating the variability calculated by ConSurf. (**C**) Extended sequence panel from the conservation calculated by ConSurf. Bold squares are highlighting the lysines and arginines localisation for all the constructs.

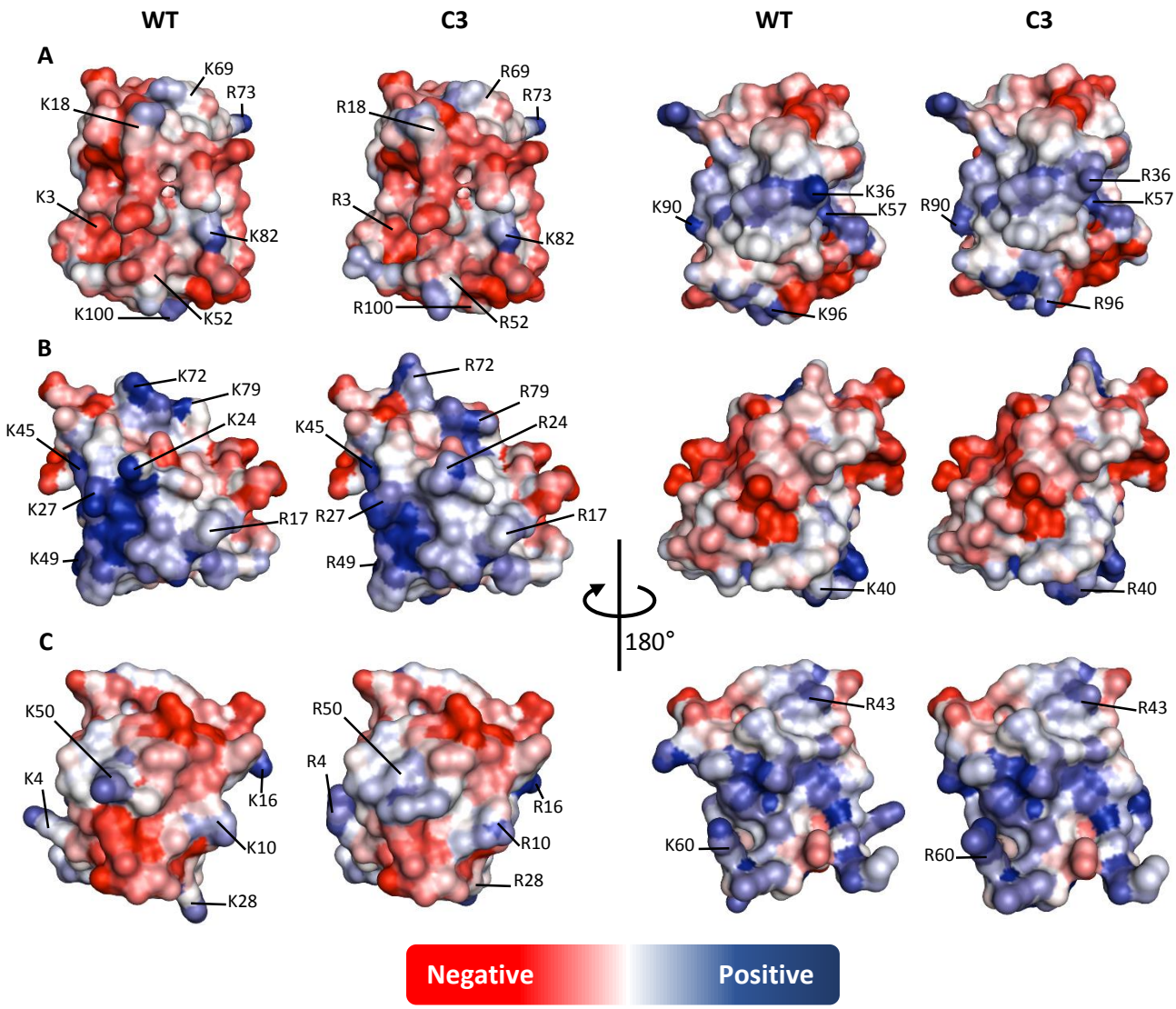


Fig. 6.2. Structural analysis of the electrostatic potential patches on protein surface. Two-sided view of the localisation of positively-charged residues (K and R) on protein surface of (**A**) thioredoxin, (**B**) HPr and (**C**) cspB. Positively-charged patches are represented by blue, non-charged patches by white and negatively charged by red colour, respectively.

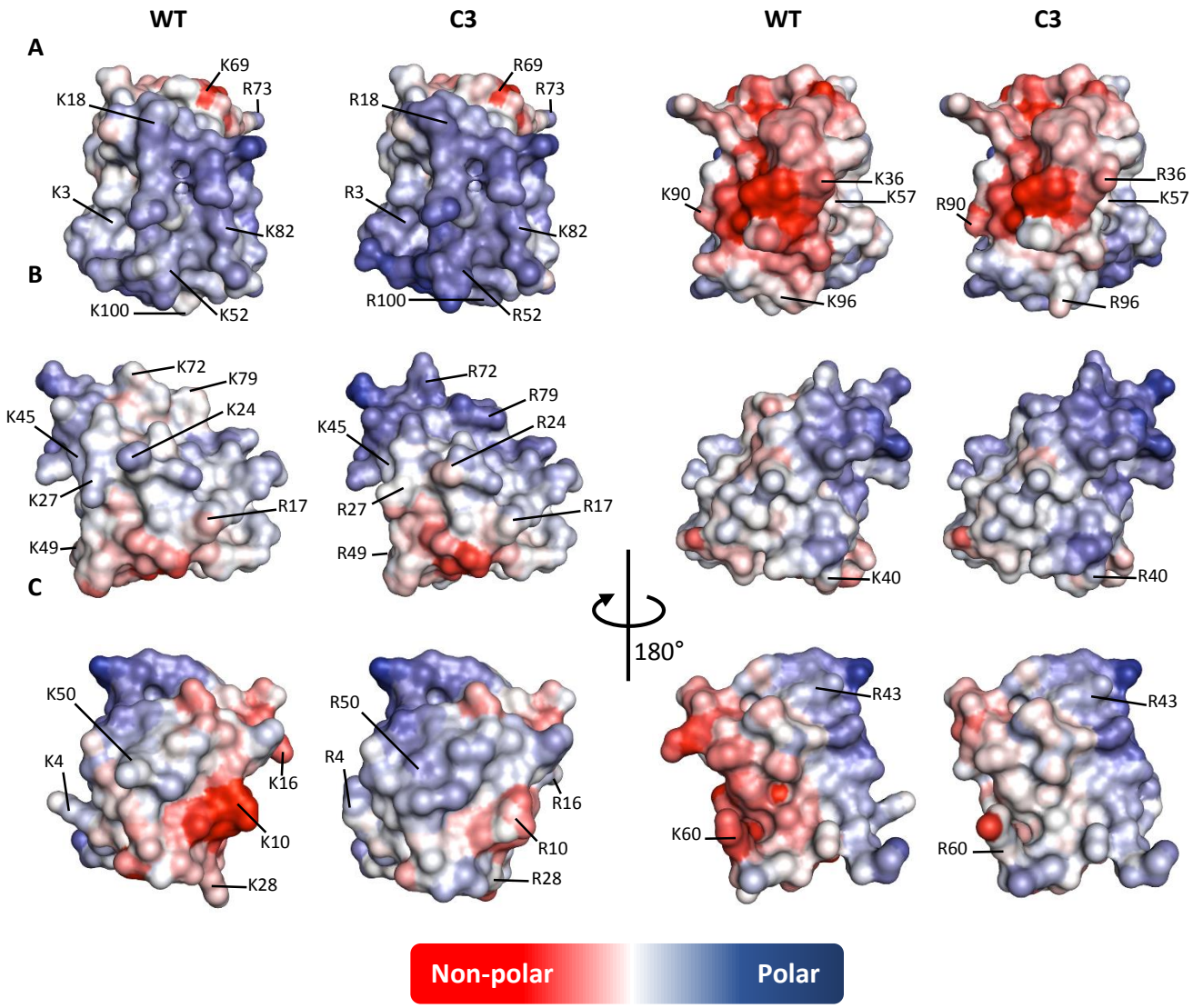


Fig. 6.3. Surface mapping of the nonpolar to polar ratio. The most non-polar (hydrophobic) areas on protein surface are highlighted as red and most polar (hydrophilic) as blue of (**A**) thioredoxin, (**B**) HPr and (**C**) cspB. A 180° view allows the full visualisation of all the K and R residues on surface.

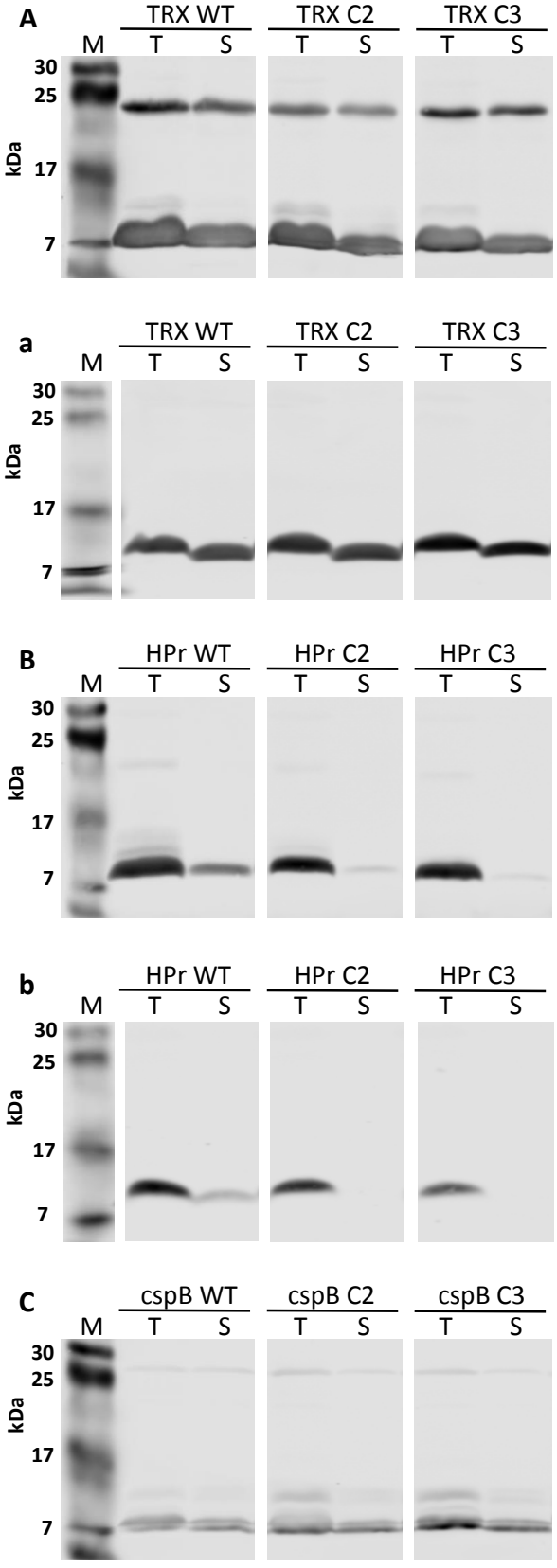
6.3.2 *Swapping lysine for arginine diminished protein solubility for HPr and cspB*

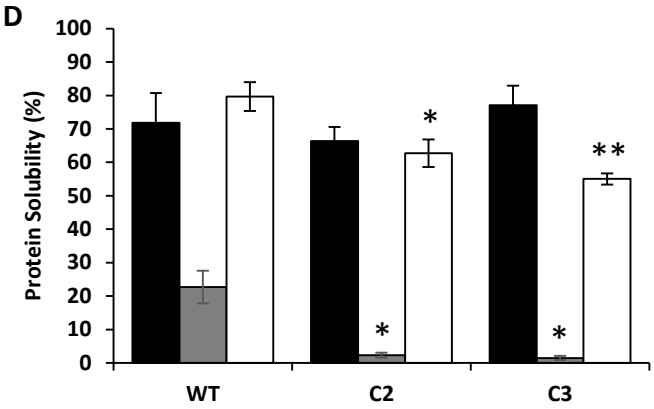
The total K:R content for thioredoxin WT is 10:1, the intermediate construct (C2) is 6:5 and 2:9 for construct 3 (C3). In the same way, cspB WT is 6:1, cspB C2 3:4 and for 0:7 cspB C3. For HPr WT is 7:1, HPr C2 4:4 and 1:7 for C3 (Table 6.2). Protein expression was carried out at the optimum growth temperature for *E. coli* of 37ºC and moderate IPTG inducer concentration (0.5 mM) and under low induction conditions (25ºC and 0.05 mM IPTG). The resulting protein solubility was calculated by densitometric difference between total and soluble fraction (Fig. 6.4). The wild-type proteins (TRX, HPr, and cspB) resulted in a more soluble version than their respective construct versions (C2 and C3) (Fig. 6.4D and F). Thioredoxin constructs did not show significant change in solubility between variants at optimum temperature. In contrast, a significant drop in solubility was achieved in all the construct variants for cspB and HPr (Table 6.2). On the other hand, TRX showed a significant drop in solubility under low induction condition (Table 6.2 and Fig. 6.4F). HPr solubility drop is consistent in both conditions (Fig. 6.4D and F). In terms of the relative total production a drop in cspB C3 compared to its WT version was observed. However, a significant increase was reached for HPr C2 and even more for C3 (Fig. 6.4E).

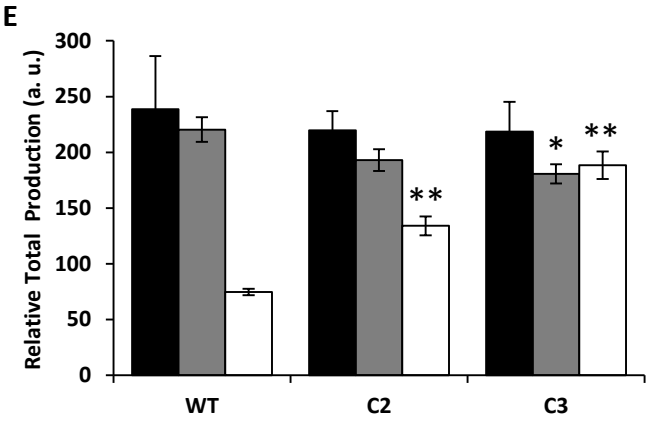
TABLE 6.2. Experimental solubility results.

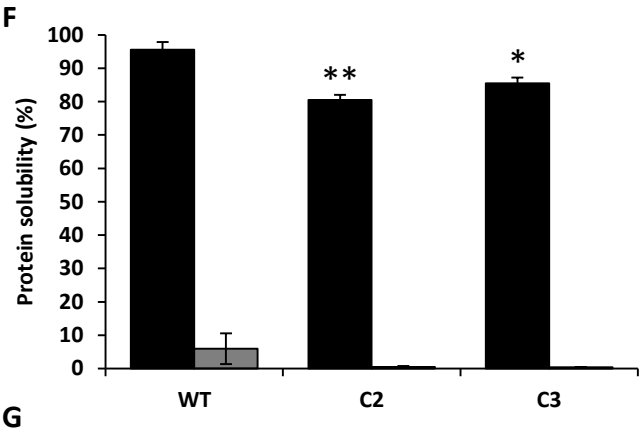
Protein (K:R ratio)	Residues	Mass (kDa)	Solubility normalised to WT	
			Optimal temperature	Sub-optimal temperature
TRX WT (10:1)	109 aa	11.8 kDa	1.00 ± 0.12	1.00 ± 0.02
TRX C2 (6:5)			0.92 ± 0.06	0.84 ± 0.02
TRX C3 (2:9)			1.07 ± 0.08	0.89 ± 0.02
HPr WT (7:1)	85 aa	9.1 kDa	1.00 ± 0.22	1.00 ± 0.78
HPr C2 (4:4)			0.10 ± 0.03	0.09 ± 0.04
HPr C3 (1:7)			0.07 ± 0.03	0.07 ± 0.01
cspB WT (6:1)	71 aa	7.7 kDa	1.00 ± 0.09	
cspB C2 (3:4)			0.82 ± 0.06	
cspB C3 (0:7)			0.72 ± 0.02	

Protein solubility analysis under optimal growth temperature and moderate inducer conditions (37ºC with 0.5 mM IPTG) and sub-optimal temperature and low inducer concentration (25ºC with 0.05 mM IPTG). Solubility profile was calculated by densitometric difference between total and soluble fraction. Triplicate biological replicates were performed for data generation and deviation represent the ± SEM.


A
TRX WT
TRX C2
TRX C3
M
T
S
kDa
30
25
17
7
a
B
HPr WT
HPr C2
HPr C3
b
C
cspB WT
cspB C2
cspB C3


D
Protein Solubility (%)
100
90
80
70
60
50
40
30
20
10
0
*
**
WT
C2
C3


E
Relative Total Production (a. u.)
300
250
200
150
100
50
0
*
**
WT
C2
C3


F
Protein solubility (%)
100
90
80
70
60
50
40
30
20
10
0
**
*
WT
C2
C3

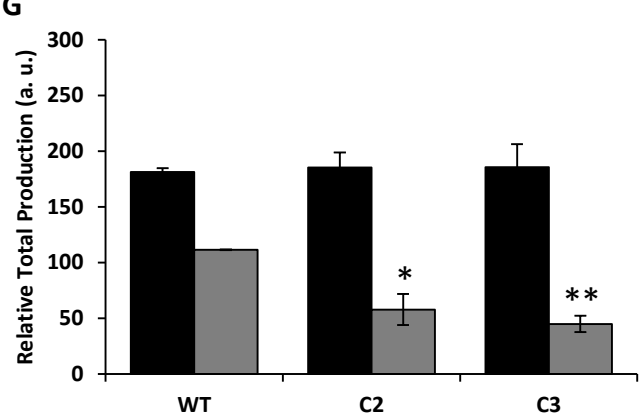

G
Relative Total Production (a. u.)
300
250
200
150
100
50
0
*
**
WT
C2
C3

Fig. 6.4. (**A-C**) Western blot of wild-type and construct protein of (**A-a**) thioredoxin, (**B-b**) HPr and (**C**) cspB. (**A-C and D-E**) Standard induction condition (37°C and induced with 0.5 mM IPTG) and (**a-b** and **F-G**) low induction condition (25°C and induced with 0.05 mM IPTG). (**D and F**) Relative protein solubility percentage between soluble and total fraction was plotted. (**E and G**) Total rHuEPO production was plotted as arbitrary units. For every BL21 (DE3) pLysS *E. coli* strain, triplicate biological replicates were performed for data generation and error bars represent the ± SE of the mean; statistically significant difference was performed using a two-sided unpaired t-test (*P < 0.05, ** P <0.01). TRX [■], HPr [■], and cspB [□]. WT, wild-type; C2, construct 2; C3, construct 3; M, prestained SDS-PAGE marker (BIO-RAD); T, Total fraction; S, Soluble fraction; a.u., arbitrary units.

6.4 Discussion

In addition to provision of fundamental understanding of the forces that maintain the fidelity of protein structure, the approach undertaken in this work to improve protein solubility for expression, and purification presents an important theme in biotechnology. One approach widely used to improve protein solubility is to decrease non-polar patches on protein surfaces. In addition, increasing negative charge has been used, and elsewhere in this thesis, we have examined the consequence of decreasing the size of positively-charged patches on solubility of expressed recombinant proteins. Here we have applied an alternative approach that involved the substitution of positive for positive residues (K to R) to improve solubility. These amino acid change mutations were located on the protein surface, and are designed to be conservative in terms of native state stability (both are positively-charged side-chains). A previous report has shown an improvement in stability when lysine to arginine mutations were made on the surface of GFP under denaturing conditions, but the thermal stability remained the same (Sokalingam et al., 2012). Furthermore, the arginine-rich variant of GFP accumulated almost entirely in the insoluble fraction. This outcome correlates with our hypothesis and our experimental findings.

Wild-type *E. coli* thioredoxin is a soluble protein when expressed. In the construct TRX C3 two highly conserved lysines remained unchanged, K57 and K82. The buried K57, is a residue essential for maintenance of redox activity (Dyson et al., 1997). The conserved K82 is important in the formation of the β4 sheet structure (Katti et al., 1990). In this study, the solubility of thioredoxin determined experimental did not change significantly between the constructs (Fig. 6.4D and Table 6.2), under our standard induction conditions. Furthermore, the relative total TRX production was the same for all the constructs (Fig. 6.4E). Low induction conditions (decreased temperature, lower IPTG challenge) were applied to the expression of

TRX WT, C2 and C3, in order to determine if the solubility properties altered. Under this sub-optimal condition TRX WT resulted in almost a 100% soluble protein. Moreover, a significant drop in solubility was achieved for the constructs (Table 6.2). An important feature of thioredoxin is its use as a stable fusion tag to improved solubility of partner proteins, correlating with the overall result that thioredoxin is particularly soluble. Lysine to arginine mutations do not significantly alter the solubility (assessed in expression) under the normal induction conditions, but do lower solubility when temperature and IPTG concentration are decreased.

Secondly, histidine-containing phosphocarrier protein (HPr) showed an insoluble profile for all the expression conditions. HPr WT solubility ranged from around 20% and solubility for HPr C2 and C3 was 2% and 1%, respectively, at the higher induction conditions. All constructs showed similar large decreases in solubility in response to the less profound induction conditions. The overall poor solubility may be due to the overexpression of this cytoplasmic protein leading to an alteration of its function *in vivo*. On the other hand, a downshift in protein production is observed for the proteins containing greater composition of arginine residues (Fig. 6.4E).

Lastly, in the cold shock-like protein cspB all lysines were substituted for arginines in construct 3. This inducible protein plays a role in the synthesis of several cold-induced proteins under sub-optimal temperatures (Graumann and Marahiel, 1997). Studies based on the homologues cspB from *Bacillus Subtilis* (Schindelin et al., 1993, Schnuchel et al., 1993) and cspA from *E. coli* (Newkirk et al., 1994, Schindelin et al., 1994), revealed the single-stranded DNA-binding action mechanism. Our experimental results suggest a significant increase (2 to 2.5-fold) in total protein expression when more arginines are displayed on cspB surface (Fig. 6.4D). This is consistent with the poor expression and degradation propensity of the homologue cspA WT at 37ºC (Brandi et al., 1996), which contains seven lysines and no arginines. It is not clear why there exists this difference in total protein amounts. With regard to the target of this

experiment, solubility changes upon mutation, results showed a significant drop in solubility for construct 2 and 3 in relation to wild-type, in line with the hypothesis.

In conclusion, the general trend for the three proteins studied in detail is for a decrease in solubility as lysines are substituted by arginines, as hypothesised from earlier computational work (Warwicker et al., 2014). The one exception was thioredoxin at higher induction levels. Since thioredoxin is used as a solubility tag, it may be that its solubility has such a high integral baseline that it is hard to perturb simply with lysine to arginine mutations. This work is a first step in studying to what degree the lysine:arginine ratio modifies solubility. Further work will take several directions. The reverse effects on solubility can be predicted, engineered and tested, with proteins of known poor solubility and relatively high arginine content (with swaps to lysine). For all proteins in these experiments, resource could be devoted to purifying protein in sufficient quantities to perform biophysical characterisations aimed at solubility analysis sway from the expression system. There is also scope for trialling different expression conditions.

Acknowledgements

We would like to thank Dr. E. McKenzie and Dr. J. Valencia for supply of the expression vector and *Escherichia coli* strains, Dr. R. Curtis for discussion, and also CONACyT for contributing PhD funds.

6.5 References

AGOSTINI, F., CIRILLO, D., LIVI, C. M., DELLI PONTI, R. & TARTAGLIA, G. G. 2014. ccSOL omics: a webserver for solubility prediction of endogenous and heterologous expression in Escherichia coli. *Bioinformatics,* 30**,** 2975-7.

AGOSTINI, F., VENDRUSCOLO, M. & TARTAGLIA, G. G. 2012. Sequence-based prediction of protein solubility. *J Mol Biol,* 421**,** 237-41.

ASHKENAZY, H., EREZ, E., MARTZ, E., PUPKO, T. & BEN-TAL, N. 2010. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res,* 38**,** W529-33.

BERG, J. M., STRYER, L. & TYMOCOZKO, J. L. 2002. The amino acid sequence of a protein determines its three-dimensional structure. *Biochemistry.* 5th ed. New York: W. H. Freeman

BRANDI, A., PIETRONI, P., GUALERZI, C. O. & PON, C. L. 1996. Post-transcriptional regulation of CspA expression in Escherichia coli. *Molecular Microbiology,* 19**,** 231-240.

CARBALLO-AMADOR, M. A., MCKENZIE, E. A., DICKSON, A. J. & WARWICKER, J. 2014a. Strategies to improve soluble expression of recombinant 6-Phosphofructo-2-Kinase/fructose-2,6-bisphosphatase (PFKFB3) in *E. coli*. *in preparation*.

CARBALLO-AMADOR, M. A., WARWICKER, J. & DICKSON, A. J. 2014b. Increasing solubility in recombinant erythropoietin through modification of surface patches. *in preparation*.

CELNIKER, G., NIMROD, G., ASHKENAZY, H., GLASER, F., MARTZ, E., MAYROSE, I., PUPKO, T. & BEN-TAL, N. 2013. ConSurf: Using Evolutionary Data to Raise Testable Hypotheses about Protein Function. *Israel Journal of Chemistry,* 53**,** 199-206.

CHAN, P., CURTIS, R. A. & WARWICKER, J. 2013. Soluble expression of proteins correlates with a lack of positively-charged surface. *Sci Rep,* 3**,** 3333.

CONCHILLO-SOLE, O., DE GROOT, N., AVILES, F., VENDRELL, J., DAURA, X. & VENTURA, S. 2007. AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. *BMC Bioinformatics,* 8**,** 65.

CONSORTIUM, T. U. 2014. Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Research,* 42**,** D191-D198.

DAVIS, G. D., ELISEE, C., NEWHAM, D. M. & HARRISON, R. G. 1999. New fusion protein systems designed to give soluble expression in Escherichia coli. *Biotechnol Bioeng,* 65**,** 382-8.

DYSON, H. J., JENG, M.-F., TENNANT, L. L., SLABY, I., LINDELL, M., CUI, D.-S., KUPRIN, S. & HOLMGREN, A. 1997. Effects of Buried Charged Groups on Cysteine Thiol Ionization and Reactivity in Escherichia coli Thioredoxin: Structural and Functional Characterization of Mutants of Asp 26 and Lys 57†. *Biochemistry,* 36**,** 2622-2636.

ETCHEGARAY, J.-P. & INOUYE, M. 1999. CspA, CspB, and CspG, Major Cold Shock Proteins ofEscherichia coli, Are Induced at Low Temperature under Conditions That Completely Block Protein Synthesis. *Journal of Bacteriology,* 181**,** 1827-1830.

FANG, Y. & FANG, J. 2013. Discrimination of soluble and aggregation-prone proteins based on sequence information. *Mol Biosyst,* 9**,** 806-11.

FERNANDEZ-ESCAMILLA, A.-M., ROUSSEAU, F., SCHYMKOWITZ, J. & SERRANO, L. 2004. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat Biotech,* 22**,** 1302-1306.

FINK, A. L. 1998. Protein aggregation: folding aggregates, inclusion bodies and amyloid. *Fold Des,* 3**,** R9-23.

FISCHER, B., SUMNER, I. & GOODENOUGH, P. 1993. Isolation, renaturation, and formation of disulfide bonds of eukaryotic proteins expressed in Escherichia coli as inclusion bodies. *Biotechnology and Bioengineering,* 41**,** 3-13.

GRAUMANN, P. & MARAHIEL, M. A. 1997. Effects of heterologous expression of CspB, the major cold shock protein of Bacillus subtillis, on protein synthesis in Escherichia coli. *Mol Gen Genet,* 253**,** 745-52.

GUEX, N. & PEITSCH, M. C. 1997. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis,* 18**,** 2714-23.

GUSTAFSSON, C., GOVINDARAJAN, S. & MINSHULL, J. 2004. Codon bias and heterologous protein expression. *Trends in Biotechnology,* 22**,** 346-353.

HATFIELD, G. W. & ROTH, D. A. 2007. Optimizing scaleup yield for protein production: Computationally Optimized DNA Assembly (CODA) and Translation Engineering™. *In:* EL-GEWELY, M. R. (ed.) *Biotechnology Annual Review.* Elsevier.

KANE, J. F. & HARTLEY, D. L. 1988. Formation of recombinant protein inclusion bodies in Escherichia coli. *Trends in Biotechnology,* 6**,** 95-101.

KATTI, S. K., LEMASTER, D. M. & EKLUND, H. 1990. Crystal structure of thioredoxin from Escherichia coli at 1.68 Å resolution. *Journal of Molecular Biology,* 212**,** 167-184.

KLUS, P., BOLOGNESI, B., AGOSTINI, F., MARCHESE, D., ZANZONI, A. & TARTAGLIA, G. G. 2014. The cleverSuite approach for protein characterization: predictions of structural properties, solubility, chaperone requirements and RNA-binding abilities. *Bioinformatics,* 30**,** 1601-8.

MARSTON, F. A. 1986. The purification of eukaryotic polypeptides synthesized in Escherichia coli. *Biochem J,* 240**,** 1-12.

MEADOW, N. D., FOX, D. K. & ROSEMAN, S. 1990. The Bacterial Phosphoenol-Pyruvate: Glycose Phosphotransferase System. *Annual Review of Biochemistry,* 59**,** 497-542.

MITRAKI, A. & KING, J. 1989. Protein Folding Intermediates and Inclusion Body Formation. *Nat Biotech,* 7**,** 690-697.

MOGK, A., MAYER, M. P. & DEUERLING, E. 2002. Mechanisms of Protein Folding: Molecular Chaperones and Their Application in Biotechnology. *ChemBioChem,* 3**,** 807-814.

NEWKIRK, K., FENG, W., JIANG, W., TEJERO, R., EMERSON, S. D., INOUYE, M. & MONTELIONE, G. T. 1994. Solution NMR structure of the major cold shock protein (CspA) from Escherichia coli: identification of a binding epitope for DNA. *Proceedings of the National Academy of Sciences of the United States of America,* 91**,** 5114-5118.

NIWA, T., YING, B. W., SAITO, K., JIN, W., TAKADA, S., UEDA, T. & TAGUCHI, H. 2009. Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of Escherichia coli proteins. *Proc Natl Acad Sci U S A,* 106**,** 4201-6.

PRILUSKY, J., FELDER, C. E., ZEEV-BEN-MORDEHAI, T., RYDBERG, E. H., MAN, O., BECKMANN, J. S., SILMAN, I. & SUSSMAN, J. L. 2005. FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics,* 21**,** 3435-8.

SAHDEV, S., KHATTAR, S. & SAINI, K. 2008. Production of active eukaryotic proteins through bacterial expression systems: a review of the existing biotechnology strategies. *Molecular and Cellular Biochemistry,* 307**,** 249-264.

SAMAK, T., GUNTER, D. & WANG, Z. 2012. Prediction of protein solubility in E. coli. *Proceedings of the 2012 IEEE 8th International Conference on E-Science (e-Science).* IEEE Computer Society.

SCHEIN, C. H. & NOTEBORN, M. H. M. 1988. Formation of Soluble Recombinant Proteins in Escherichia Coli is Favored by Lower Growth Temperature. *Nat Biotech,* 6**,** 291-294.

SCHINDELIN, H., JIANG, W., INOUYE, M. & HEINEMANN, U. 1994. Crystal structure of CspA, the major cold shock protein of Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America,* 91**,** 5119-5123.

SCHINDELIN, H., MARAHIEL, M. A. & HEINEMANN, U. 1993. Universal nucleic acid-binding domain revealed by crystal structure of the B. subtilis major cold-shock protein. *Nature,* 364**,** 164-168.

SCHNUCHEL, A., WILTSCHECK, R., CZISCH, M., HERRLER, M., WLLLLMSKY, G., GRAUMANN, P., MARAHIEL, M. A. & HOLAK, T. A. 1993. Structure in solution of the major cold-shock protein from Bacillus subtilis. *Nature,* 364**,** 169-171.

SCHRÖDINGER, L. L. C. 2010. The PyMOL Molecular Graphics System, Version 1.3r1.

SOKALINGAM, S., RAGHUNATHAN, G., SOUNDRARAJAN, N. & LEE, S.-G. 2012. A Study on the Effect of Surface Lysine to Arginine Mutagenesis on Protein Stability and Structure Using Green Fluorescent Protein. *PLoS ONE,* 7**,** e40410.

TARTAGLIA, G. G. & VENDRUSCOLO, M. 2008. The Zyggregator method for predicting protein aggregation propensities. *Chemical Society Reviews,* 37**,** 1395-1401.

THELANDER, L. 1967. Thioredoxin Reductase: CHARACTERIZATION OF A HOMOGENEOUS PREPARATION FROM ESCHERICHIA COLI B. *Journal of Biological Chemistry,* 242**,** 852-859.

WALSH, I., SENO, F., TOSATTO, S. C. E. & TROVATO, A. 2014. PASTA 2.0: an improved server for protein aggregation prediction. *Nucleic Acids Research,* 42**,** W301-W307.

WARWICKER, J., CHARONIS, S. & CURTIS, R. A. 2014. Lysine and Arginine Content of Proteins: Computational Analysis Suggests a New Tool for Solubility Design. *Molecular Pharmaceutics,* 11**,** 294-303.

WATERHOUSE, A. M., PROCTER, J. B., MARTIN, D. M., CLAMP, M. & BARTON, G. J. 2009. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics,* 25**,** 1189-91.

WEICKERT, M. J., DOHERTY, D. H., BEST, E. A. & OLINS, P. O. 1996. Optimization of heterologous protein production in Escherichia coli. *Current Opinion in Biotechnology,* 7**,** 494-499.

YASUKAWA, T., KANEI-ISHII, C., MAEKAWA, T., FUJIMOTO, J., YAMAMOTO, T. & ISHII, S. 1995. Increase of solubility of foreign proteins in Escherichia coli by coproduction of the bacterial thioredoxin. *J Biol Chem,* 270**,** 25328-31.

Chapter 7

Concluding remarks

7.1 Overall discussion

The rapid growth of the world population is accompanied by demands for drug development in the clinical practice. To fulfil these needs several strategies have been developed ranging from genetic and cellular engineering to protein design. Protein-based drugs are one of the most profitable sectors in the therapeutic market (Pavlou and Reichert, 2004, Walsh, 2014). However, the demand of proteins is not exclusively for clinical use, it is also extends to various applications such as biochemical analysis or structural studies (Esposito and Chatterjee, 2006). The enterobacterium *Escherichia coli* is an outstanding expression system, since it generally yields high amounts of recombinant proteins and offers cost-efficient production (Hirose et al., 2011). Nevertheless, the overexpression frequently leads to degradation of misfolded protein or aggregation into inclusion bodies (Hoffmann and Rinas, 2004). To overcome these adverse mechanisms, several strategies have been proposed, including use of different *E. coli* strains, protein expression at low temperatures, different cultivation strategies, co-expression of molecular chaperones, fusion of desired proteins with solubility enhancing tags and "rational" site-directed mutagenesis (Sørensen and Mortensen, 2005b). In addition, over the past decades several computational approaches have been developed to understand and address protein solubility and aggregation issues (Wilkinson and Harrison, 1991, Hwang and Park, 2008, Chan et al., 2013, Chang et al., 2014, Warwicker et al., 2014). Unfortunately, there is no universal

strategy to overcome the problems of protein solubility and aggregation upon heterologous expression (Sørensen and Mortensen, 2005a).

Some of the approaches mentioned above have been addressed in the preceding chapters in this thesis. The results have been presented as four chapters in a style suitable for the intended journal of submission. The overall aim was to provide the first specific experimental tests for computational approaches developed in our group, through protein surface charge patch engineering in relation in two expression systems: *Escherichia coli* and HEK 293-EBNA cells. In the following sections, the key questions based on the aims in this thesis are discussed.

7.1.1 What were the individual value of the three computational approaches to the prediction of solubility of PFKFB3?

To our knowledge, this is the first study of any PFKFB isoform based on predicted calculation to enhance protein solubility and stability. PFKFB3 in previous expression studies mostly accumulated in inclusion bodies (McKenzie E.A., unpublished data). Based on those preliminary experiments we investigated three different strategies to improve the solubility of expression in three *E. coli* strains (Chapter 3, paper 1). These strategies comprised rational protein design utilising the predictive algorithm developed in our group (Chan et al., 2013) to (i) diminish non-polar areas, (ii) decrease positively-charged patches and (iii) introduce a charged helical cap into thermo-flexible areas.

PFKFB3 solubility was assessed by immunoblotting of total and soluble fractions. The expression of human PFKFB3 in three different cellular environments (standard and controlled

expression system [BL21 (DE3) pLysS], rare codon recognition [BL21-CodonPlus], and disulphide bond formation/chaperone-folding [SHuffle] system) demonstrated variability among the solubility results, between mutants and expression systems. Among the strategies explored, only diminishing positive patch size offers a consistent solubility improvement in all expression systems for mutant M2. In contrast, for the second mutant (M4) aimed at this strategy, results depend on expression system, such as SHuffle strain.

According to our findings, the widely-described strategy of hydrophobicity plotting of protein surfaces did not offer an answer to improve PFKFB3 solubility. However, the oxidative environment and the presence of a periplasmic chaperone in the cytoplasm, contribute to a more soluble version of M1 using the SHuffle system. The inconsistency in solubility results suggests that the non-polar area of M1 is not the most important factor in aggregate formation. Additionally, the predicted decrease in local flexibility did not improve solubility among *E. coli* strains, rather mutant M3 (aim to diminished local flexibility) consistently diminished solubility.

PFKFB3 exhibits a relatively large maximal positive patch in comparison with a threshold derived in a recent study (Chan et al., 2013). Based on the experimental results, this feature stood out of the three computational strategies, despite the relatively small increase in solubility. This algorithm provided the foundations for the next two experimental chapters.

7.1.2 Does positively-charged patches size influence soluble expression of rHuEPO in *E. coli*?

Trevino and co-workers showed that aspartic acid, glutamic acid, and serine contribute more favourably to RNase Sa solubility among the 20 amino acid residues (Trevino et al., 2007). Also, several groups have demonstrated a favourable influence of negatively-charged surface amino acids on solubility (Dale et al., 1994, Zhang et al., 1997, Fan et al., 2004). One possible explanation is that carboxyl groups of aspartic and glutamic acid bind to water more strongly than do amino and guanidine groups of arginine and lysine (Kuntz, 1971, Collins and Washabaugh, 1985, Collins, 1997, Kramer et al., 2012, Chong and Ham, 2014). In addition, our group recently demonstrated a correlation between positive charge patches and insolubility in a cell-free expression system database (Chan et al., 2013). An algorithm arose from this publication, which computes structured-based parameters, including the maximal size of non-charged and positively charged patch and their multiplicative combination, versus thresholds calculated from Niwa dataset of experimental solubilities (Niwa et al., 2009).

A set of rHuEPO proteins was generated by adjusting positively-charged patches with positively- or negatively-charged amino acid mutations (Chapter 4, paper 2). Experimental solubility was determined upon expression in two *E. coli* strains (Bl21 (DE3) pLysS and SHuffle) at lower induction conditions (sub-optimal temperature and low inducer concentration). In both strains, an engineered larger positively-charged patch successfully increased IB formation in rHuEPO E13K, as predicted. In addition, we detected a significant increase in solubility for the less positively-charged patch size variants in the SHuffle strain and rHuEPO R150D in BL21 (DE3) pLysS. The results with rHuEPO, for positive charge patch engineered larger and smaller, are therefore encouraging.

We have successfully tested the hypothesis that positive charge patches can influence protein solubility in expression, using the example of rHuEPO. This study sets the starting point for further investigation, including a study of the rHuEPO variants in a mammalian expression system.

7.1.3 Are the *E. coli*-derived rHuEPO aggregation results translatable to the secretory environment in HEK 293-EBNA cells?

Proteins are one of the most profitable products among the current biopharmaceutical manufacturing industry (Pavlou and Reichert, 2004). To date, more than half of all recombinant therapeutic proteins have been produced in mammalian cells, mainly due to the high similarity of the final product to human protein (PTM) structures (Matasci et al., 2008). Also, most of the commercial therapeutics are secreted proteins (e.g. hormones, interferon, monoclonal antibodies) (Peng and Fussenegger, 2009). Therefore, research groups have been working to maximise the translational or secretory capacity of mammalian host cells (Barnes and Dickson, 2006). A part of the solution is likely to depend on progress in the field of synthetic biology, which includes increasing stability and solubility by protein design (Caravella and Lugovskoy, 2010).

We investigated the consequences of expressing a synthetic redesigned set of rHuEPO variants with altered surface charge (Section 4) in HEK 293-EBNA cells (Chapter 5, paper 3). Diminishing the largest positively-charged patch (rHuEPO F48D, R150D and F48D/R150D) favoured production of secreted product compared to rHuEPO WT. In contrast, a variant with a higher positively-charged patch (rHuEPO E13K) resulted in the lowest amount of secreted protein among all the expression variants.

A correlation between poor product yield and the extent of positive charge on protein surface patches was observed. This approach could offer a potential tool for the rational design of proteins with improved stability and secretion along the secretory pathway.

7.1.4 Does the lysine:arginine content influence protein solubility?

Since there is no universal approach or technology that solves protein aggregation, more detailed and extensive observations need to be made. One approach widely used to prevent protein aggregation is to decrease non-polar patches on protein surfaces. In addition, we have investigated the consequence of decreasing the size of positively-charged patches on solubility of expressed recombinant proteins in the preceding chapters in this thesis (Chapter 3-5). Recently, when our group observed positively-charged patches features, it was found that the sequence-based property of lysine *versus* arginine content discriminates between the least and more soluble subsets of *E. coli* protein (Warwicker et al., 2014). A greater lysine:arginine content was associated with the more soluble subset.

We investigated the solubility effect of three high soluble *E. coli* proteins with varying degree of lysine substitution by arginine (Chapter 6, paper 4). These experiments are predicted to decrease the solubility of the variants, according to our computational calculations. These proteins were: thioredoxin (TRX), cold shock-like protein cspB, and the histidine-containing phosphocarrier protein (HPr). Our findings revealed a significant decrease in solubility for cspB and HPr, which is more evident in variants with low or null lysine content. Under the same expression conditions, TRX did not change significantly. However, TRX expression under low induction conditions (low temperature and IPTG inducer) significant falls in solubility were observed for the variants compared to WT.

This computational and experimental approach is a first step in studying to what extent the lysine:arginine ratio modifies solubility. The encouraging results obtained in this thesis, with computation pointing the way towards solubility increase, will form the basis for further studies. In particular, biophysical characterisation of protein-protein interactions for purified

proteins, and increasing the range of proteins and variants being tested, will be necessary for more full evaluation of the underlying molecular mechanisms. Even at this stage though, it is clear that the current work can contribute to rationalisation of means to improve yields in recombinant protein technologies.

7.2 Future vision

The study presented in this thesis denotes the first experimental work derived from two novel computational algorithms recently developed in our group. The experimental exploration of the preceding chapters encompasses tests on a relatively small set of proteins undertaken to provide experimental validation of computational approaches. While answering many initial questions the results presented have the potential for extension and, in addition, generate new hypotheses for further investigation.

In Chapter 3, PFKFB3 was chosen as it represented a challenging "prone-to-aggregation" protein from our experimental and preliminary computational studies. PFKFB3 wild-type was predicted to be an insoluble protein upon heterologous expression in *E. coli*. The numerical output of this prediction was relatively high, and this was the case for all PFKFB3 variants examined in this Chapter. These predictions were reflected in the relatively small increment in experimental solubility, when the largest positively-charged patch size was modified. However, it would be possible to explore multiple mutations at different positions of the patch in order to decrease the overall patch size. This has the potential to generate a more favourable solubility prediction, a prediction amenable to experimental testing. In addition, the combinatorial strategy of increased polarity (mutant M1; W13Y, V14N, V16K) with diminished positively-charged patch (mutant M4; R427D) in the SHuffle system would be

predicted to generate a product of greater solubility based on our results. Also, PFKFB3 illustrates an important point about the need for completeness of crystal structure information for the veracity of computational-based predictions. Two small areas (K29 to T31 and E446 to P452) and most of the C-terminal section (from V461) is not available from the PFKFB3 crystal structure (Kim et al., 2006) and such aspects may compromise the output from computational predictions. Further modelling or structural studies may offer a better understanding and prediction accuracy of our algorithms. We have investigated the outcome of computational-directed protein engineering strategies in relation to *in vivo* solubility when expressed in *E. coli*. No improvement in solubility was achieved but purification of the mutated variants would allow further analyses of structural relationships to, for example, enzymatic activity, protein stability, and protein-protein interactions. Simple colorimetric enzyme-coupled kinetic assays have been developed for PFKFB3 activity (Kim et al., 2007, Clem et al., 2008, Seo et al., 2011). A variety of biophysical methods could be applied for stability measurements, including light scattering and fluorescence technology, with measurements of unfolding transition temperature (Tm), aggregation onset temperature (Tagg) and rates of aggregation (Goldberg et al., 2011, Avacta, 2013). Furthermore, structural studies such as size exclusion chromatography could offer evidence of an alteration of dimerisation among PFKFB3 variants. Mutant M3 (a variant with diminished local flexibility) would present an intriguing study.

Despite the promising findings in Chapter 4, these results were generated from a single protein, human erythropoietin. This protein is a good model to test the computational predictions, as it requires complex PTMs (e.g. glycosylation and disulphide bond formation), encompasses a relative low protein mass and is a biopharmaceutical with high economic expectations for biosimilar products and analogues (biobetters) (Jelkmann, 2013, Mikhail and Farouk, 2013). The results revealed a correlation between the positively-charged patch and soluble expression in *E. coli*, with the strongest correlation arising from studies with the

SHuffle strain (a strain with an oxidising cytoplasmic environment (Berkmen, 2012)). It would be intriguing to further explore how the rHuEPO variants generated in this thesis would express when targeted to the periplasmic oxidative environment (Depuydt et al., 2009). The enhanced solubility of expression in the oxidative environment of SHuffle cytoplasm (which lacks thioredoxin reductase [*trxB*] and glutathione reductase [*gor*] but contains DsbC disulphide bond isomerase (Lobstein et al., 2012)), suggests that disulphide bridge formation (C7-C161 and C29-C33) plays a significant role in translation of computational prediction to experimental outcome. It would be possible to identify the presence of disulphide linkages, in support of this hypothesis, by MALDI-TOF MS analysis. The application of mass-spectrometry would also be valuable in determination of the molecular reason for the molecular mass shift observed for the F48D mutation of rHuEPO (Fig. 4.4). Whilst the chemical modification (at a specific residue) may not present functional consequences in this case, it may provide valuable insight to understanding the molecular consequences for protein engineering strategies in general.

Increasing the soluble rHuEPO repertoire by study of other mutants (e.g. mutant G158E in the rHuEPO G09 from MedImmune (Buchanan et al., 2012)) may give a robust validation for this globular hormone. Based on the model developed here for rHuEPO, other proteins (e.g. from the cytokine four-α-helix bundle family [interleukins and interferons]) could be added to extend the validation of the predictive computational approach. Towards an overall aim of production of high soluble protein yields, different expression conditions may be applied to this set of rHuEPO variants (and other proteins). These may include optimum *E. coli* growth temperature, codon optimisation or use of further specialised *E. coli* strains (i.e. BL21-CodonPlus or Rosetta strains), using solubility-enhancing fusion tags, and co-expression of chaperones.

In Chapter 5, rHuEPO E13K exhibited poor secretion and this was hypothesised to arise from the extent of positive charge on surface patches. Before any general correlation can be made between positively-charged patch size on rHuEPO and production through the secretory pathway of HEK 293-EBNA cells there is need for several further studies. The rHuEPO E13K variant alone offers further exploration at cellular and molecular level to understand limitations in the formation of secreted protein. These studies may involve the analysis and detection of the endoplasmic reticulum (ER) regulator chaperones such as BiP (or binding immunoglobulin protein; GRP78), XBP1 (X-box binding protein 1) or the pro-apoptotic C/EBP homologous protein (CHOP or GADD153). These target proteins may indicate the presence of cellular stress upon folding and PTMs of rHuEPO variants in the ER (Samali et al., 2010). This is relevant since rHuEPO requires PTMs such as N-linked glycosylation, which is processed in the ER (Chakrabarti et al., 2011_ENREF_39). Also, rHuEPO degradation rate could be measured using flow cytometry with a GFP reporter (Yewdell et al., 2011). Furthermore, rHuEPO protein localisation within compartments of the secretory pathway could be undertaken by confocal microscopy. Purification of rHuEPO WT and variants (secreted and intracellular) by nickel immobilized metal chelate affinity chromatography and size exclusion chromatography will allow protein quantification and characterisation. Biomolecular and biophysical analyses would offer molecular profiling of the extent to which post-translational modifications (events key for the guided passage of rHuEPO to the extracellular environment) have occurred and, potentially, provide indicators of events associated with the effectiveness of secretion. Whilst the focus for the current studies has been on the ability to increase the amount of EPO protein expressed from variants, less protein may not mean less functional protein. Buchanan et al., (2012), generated mutant rHuEPO G09 (with altered posQ) and rHuEPO G09 had greater specific activity of 8.5-fold than rHuEPO WT. Mutants generated in the current study can be

tested for biological activity by use of an *in vitro* proliferation bioassay of TF-1 cells (Kitamura et al., 1989).

The favourable results of altering the lysine and arginine content in Chapter 6, suggests a series of further studies in order to gain a better understanding of this hypothesis. We investigated decrease in solubility of highly soluble *E. coli* proteins by changing lysines to arginines. Therefore, engineering low soluble *E. coli* proteins to increase solubility (arginines to lysines) is essential for the whole understanding. In addition, the modification of proteins of mid-range (eSOL database index) solubility, whether increasing or decreasing lysine content, would be of interest to expand the experimental spectrum. A further application is to involve a larger number of *E. coli* proteins to validate this computational strategy. Moreover, further exploration of non-*E.coli* source proteins, such as therapeutic or structural proteins of interest will be useful. Biophysical studies upon protein purification could also play an important role in providing feedback for the computational algorithm.

To summarise, all the experimental chapters could lead to further purification and characterisation. This would be facilitated due to the presence of a thrombin-cleavable amino-terminal 6x His tag in all the constructs, opening up further biophysical and activity studies. Also, more variants or number of proteins will be useful for validation of the different approaches developed in this thesis. In addition, the use of different expression conditions such as environmental (e.g. nutrients or temperature), inducer concentration or time post-induction could help to investigate the hypotheses. The use of different expression systems could lead to better understanding of the computational approaches, e.g. using a cell-free system which offers a relative fast and extensive screening.

Chapter 8

References

AGOSTINI, F., CIRILLO, D., LIVI, C. M., DELLI PONTI, R. & TARTAGLIA, G. G. 2014. ccSOL omics: a webserver for solubility prediction of endogenous and heterologous expression in Escherichia coli. *Bioinformatics,* 30**,** 2975-7.

AGOSTINI, F., VENDRUSCOLO, M. & TARTAGLIA, G. G. 2012. Sequence-based prediction of protein solubility. *J Mol Biol,* 421**,** 237-41.

AHN, J. H., KEUM, J. W. & KIM, D. M. 2011. Expression screening of fusion partners from an E. coli genome for soluble expression of recombinant proteins in a cell-free protein synthesis system. *PLoS One,* 6**,** e26875.

AILOR, E. & BETENBAUGH, M. J. 1998. Overexpression of a cytosolic chaperone to improve solubility and secretion of a recombinant IgG protein in insect cells. *Biotechnology and Bioengineering,* 58**,** 196-203.

AIYAR, A., TYREE, C. & SUGDEN, B. 1998. The plasmid replicon of EBV consists of multiple cis-acting elements that facilitate DNA synthesis by the cell and a viral maintenance element. *The EMBO Journal,* 17**,** 6394-6403.

ALLEN, G. A., PERSSON, E., CAMPBELL, R. A., EZBAN, M., HEDNER, U. & WOLBERG, A. S. 2007. A variant of recombinant factor VIIa with enhanced procoagulant and antifibrinolytic activities in an in vitro model of hemophilia. *Arterioscler Thromb Vasc Biol,* 27**,** 683-9.

ANDYA, J. D., HSU, C. C. & SHIRE, S. J. 2003. Mechanisms of aggregate formation and carbohydrate excipient stabilization of lyophilized humanized monoclonal antibody formulations. *AAPS PharmSci,* 5**,** E10.

ANFINSEN, C. B. 1973. Principles that Govern the Folding of Protein Chains. *Science,* 181**,** 223-230.

ARICESCU, A. R., ASSENBERG, R., BILL, R. M., BUSSO, D., CHANG, V. T., DAVIS, S. J., DUBROVSKY, A., GUSTAFSSON, L., HEDFALK, K., HEINEMANN, U., JONES, I. M., KSIAZEK, D., LANG, C., MASKOS, K., MESSERSCHMIDT, A., MACIEIRA, S., PELEG, Y., PERRAKIS, A., POTERSZMAN, A., SCHNEIDER, G., SIXMA, T. K., SUSSMAN, J. L., SUTTON, G., TARBOUREICH, N., ZEEV-BEN-MORDEHAI, T. & JONES, E. Y. 2006. Eukaryotic expression: developments for structural proteomics. *Acta Crystallographica Section D,* 62**,** 1114-1124.

ARYA, R., BHATTACHARYA, A. & SAINI, K. S. 2008. Dictyostelium discoideum—a promising expression system for the production of eukaryotic proteins. *The FASEB Journal,* 22**,** 4055-4066.

ASANO, Y., DADASHIPOUR, M., YAMAZAKI, M., DOI, N. & KOMEDA, H. 2011. Functional expression of a plant hydroxynitrile lyase in Escherichia coli by directed evolution: creation and characterization of highly in vivo soluble mutants. *Protein Eng Des Sel,* 24**,** 607-16.

ASHKENAZY, H., EREZ, E., MARTZ, E., PUPKO, T. & BEN-TAL, N. 2010. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res,* 38**,** W529-33.

AVACTA 2013. Predicting Monoclonal Antibody Stability in Different Formulations Using Optim 2. Application Note. Avacta Analytical, UK.

BAGBY, S., TONG, K. I. & IKURA, M. 2001. [2] - Optimization of Protein Solubility and Stability for Protein Nuclear Magnetic Resonance. *In:* THOMAS L. JAMES, V. D. & ULI, S. (eds.) *Methods in Enzymology*. Academic Press.

BANEYX, F. 1999. Recombinant protein expression in Escherichia coli. *Current Opinion in Biotechnology,* 10**,** 411-421.

BANEYX, F. & MUJACIC, M. 2004. Recombinant protein folding and misfolding in Escherichia coli. *Nat Biotech,* 22**,** 1399-1408.

BANKS, D. D. 2011. The Effect of Glycosylation on the Folding Kinetics of Erythropoietin. *Journal of Molecular Biology,* 412**,** 536-550.

BARDERAS, R., DESMET, J., TIMMERMAN, P., MELOEN, R. & CASAL, J. I. 2008. Affinity maturation of antibodies assisted by in silico modeling. *Proc Natl Acad Sci U S A,* 105**,** 9029-34.

BARNES, L. M. & DICKSON, A. J. 2006. Mammalian cell factories for efficient and stable protein expression. *Current Opinion in Biotechnology,* 17**,** 381-386.

BAZAN, J. F., FLETTERICK, R. J. & PILKIS, S. J. 1989. Evolution of a bifunctional enzyme: 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase. *Proc Natl Acad Sci U S A,* 86**,** 9642-6.

BECKMANN, R., MIZZEN, L. & WELCH, W. 1990. Interaction of Hsp 70 with newly synthesized proteins: implications for protein folding and assembly. *Science,* 248**,** 850-854.

BERG, J. M., STRYER, L. & TYMOCOZKO, J. L. 2002. The amino acid sequence of a protein determines its three-dimensional structure. *Biochemistry.* 5th ed. New York: W. H. Freeman

BERKMEN, M. 2012. Production of disulfide-bonded proteins in Escherichia coli. *Protein Expression and Purification,* 82**,** 240-251.

BHOPALE, G. M. & NANDA, R. K. 2005. Recombinant DNA expression products for human therapeutic use. *Current Science,* 89**,** 614-622.

BLOM, N., SICHERITZ-PONTÉN, T., GUPTA, R., GAMMELTOFT, S. & BRUNAK, S. 2004. Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *PROTEOMICS,* 4**,** 1633-1649.

BONDOS, S. E. & BICKNELL, A. 2003. Detection and prevention of protein aggregation before, during, and after purification. *Anal Biochem,* 316**,** 223-31.

BRANDI, A., PIETRONI, P., GUALERZI, C. O. & PON, C. L. 1996. Post-transcriptional regulation of CspA expression in Escherichia coli. *Molecular Microbiology,* 19**,** 231-240.

BRINKS, V., HAWE, A., BASMELEH, A. H., JOACHIN-RODRIGUEZ, L., HASELBERG, R., SOMSEN, G. W., JISKOOT, W. & SCHELLEKENS, H. 2011. Quality of original and biosimilar epoetin products. *Pharm Res,* 28**,** 386-93.

BRONDYK, W. H. 2009. Chapter 11 Selecting an Appropriate Method for Expressing a Recombinant Protein. *In:* RICHARD, R. B. & MURRAY, P. D. (eds.) *Methods in Enzymology*. Academic Press.

BROOKE, D. G., VAN DAM, E. M., WATTS, C. K., KHOURY, A., DZIADEK, M. A., BROOKS, H., GRAHAM, L. J., FLANAGAN, J. U. & DENNY, W. A. 2014. Targeting the Warburg Effect in cancer; relationships for 2-arylpyridazinones as inhibitors of the key glycolytic enzyme 6-phosphofructo-2-kinase/2,6-bisphosphatase 3 (PFKFB3). *Bioorg Med Chem,* 22**,** 1029-39.

BUCHANAN, A., FERRARO, F., RUST, S., SRIDHARAN, S., FRANKS, R., DEAN, G., MCCOURT, M., JERMUTUS, L. & MINTER, R. 2012. Improved drug-like properties of therapeutic proteins by directed evolution. *Protein Eng Des Sel,* 25**,** 631-8.

CARAVELLA, J. & LUGOVSKOY, A. 2010. Design of next-generation protein therapeutics. *Current Opinion in Chemical Biology,* 14**,** 520-528.

CARAVELLA, J. A., WANG, D., GLASER, S. M. & LUGOVSKOY, A. 2010. Structure-Guided Design of Antibodies. *Curr Comput Aided Drug Des*.

CARBALLO-AMADOR, M. A., MCKENZIE, E. A., DICKSON, A. J. & WARWICKER, J. 2014a. Strategies to improve soluble expression of recombinant 6-Phosphofructo-2-Kinase/fructose-2,6-bisphosphatase (PFKFB3) in *E. coli*. *in preparation*.

CARBALLO-AMADOR, M. A., WARWICKER, J. & DICKSON, A. J. 2014b. Increasing solubility in recombinant erythropoietin through modification of surface patches. *in preparation*.

CARPENTER, J. F., RANDOLPH, T. W., JISKOOT, W., CROMMELIN, D. J. A., MIDDAUGH, C. R. & WINTER, G. 2010. Potential inaccurate quantitation and sizing of protein aggregates by size exclusion chromatography: Essential need to use orthogonal methods to assure the quality of therapeutic protein products. *Journal of Pharmaceutical Sciences,* 99**,** 2200-2208.

CAVALIER, M. C., KIM, S. G., NEAU, D. & LEE, Y. H. 2012. Molecular basis of the fructose-2,6-bisphosphatase reaction of PFKFB3: transition state and the C-terminal function. *Proteins,* 80**,** 1143-53.

CELNIKER, G., NIMROD, G., ASHKENAZY, H., GLASER, F., MARTZ, E., MAYROSE, I., PUPKO, T. & BEN-TAL, N. 2013. ConSurf: Using Evolutionary Data to Raise Testable Hypotheses about Protein Function. *Israel Journal of Chemistry,* 53**,** 199-206.

CHAKRABARTI, A., CHEN, A. W. & VARNER, J. D. 2011. A review of the mammalian unfolded protein response. *Biotechnology and Bioengineering,* 108**,** 2777-2793.

CHAN, P., CURTIS, R. A. & WARWICKER, J. 2013. Soluble expression of proteins correlates with a lack of positively-charged surface. *Sci Rep,* 3**,** 3333.

CHANG, C. C. H., SONG, J., TEY, B. T. & RAMANAN, R. N. 2014. Bioinformatics approaches for improved recombinant protein production in Escherichia coli: protein solubility prediction. *Briefings in Bioinformatics,* 15**,** 953-962.

CHEETHAM, J. C., SMITH, D. M., AOKI, K. H., STEVENSON, J. L., HOEFFEL, T. J., SYED, R. S., EGRIE, J. & HARVEY, T. S. 1998. NMR structure of human erythropoietin and a comparison with its receptor bound conformation. *Nat Struct Biol,* 5**,** 861-6.

CHENNAMSETTY, N., VOYNOV, V., KAYSER, V., HELK, B. & TROUT, B. L. 2009. Design of therapeutic proteins with enhanced stability. *Proc Natl Acad Sci U S A,* 106**,** 11937-42.

CHENNAMSETTY, N., VOYNOV, V., KAYSER, V., HELK, B. & TROUT, B. L. 2011. Prediction of protein binding regions. *Proteins,* 79**,** 888-97.

CHESNEY, J., MITCHELL, R., BENIGNI, F., BACHER, M., SPIEGEL, L., AL-ABED, Y., HAN, J. H., METZ, C. & BUCALA, R. 1999. An inducible gene product for 6-phosphofructo-2-kinase with an AU-rich instability element: Role in tumor cell glycolysis and the Warburg effect. *Proceedings of the National Academy of Sciences,* 96**,** 3047-3052.

CHESSHYRE, J. & HIPKISS, A. 1989. Low temperatures stabilize interferon α-2 against proteolysis in Methylophilus methylotrophus and Escherichia coli. *Applied Microbiology and Biotechnology,* 31**,** 158-162.

CHONG, S. H. & HAM, S. 2014. Interaction with the surrounding water plays a key role in determining the aggregation propensity of proteins. *Angew Chem Int Ed Engl,* 53**,** 3961-4.

CHUNG, C., LIU, J., EMILI, A. & FREY, B. J. 2011. Computational refinement of post-translational modifications predicted from tandem mass spectrometry. *Bioinformatics,* 27**,** 797-806.

CLARK, L. A., BORIACK-SJODIN, P. A., ELDREDGE, J., FITCH, C., FRIEDMAN, B., HANF, K. J., JARPE, M., LIPAROTO, S. F., LI, Y., LUGOVSKOY, A., MILLER, S., RUSHE, M., SHERMAN, W., SIMON, K. & VAN VLIJMEN, H. 2006. Affinity enhancement of an in vivo matured therapeutic antibody using structure-based computational design. *Protein Sci,* 15**,** 949-60.

CLEM, B., TELANG, S., CLEM, A., YALCIN, A., MEIER, J., SIMMONS, A., RASKU, M. A., ARUMUGAM, S., DEAN, W. L., EATON, J., LANE, A., TRENT, J. O. & CHESNEY, J. 2008. Small-molecule inhibition of 6-phosphofructo-2-kinase activity suppresses glycolytic flux and tumor growth. *Mol Cancer Ther,* 7**,** 110-20.

CLEM, B. F., O'NEAL, J., TAPOLSKY, G., CLEM, A. L., IMBERT-FERNANDEZ, Y., KERR, D. A., 2ND, KLARER, A. C., REDMAN, R., MILLER, D. M., TRENT, J. O., TELANG, S. & CHESNEY, J. 2013. Targeting 6-phosphofructo-2-kinase (PFKFB3) as a therapeutic strategy against cancer. *Mol Cancer Ther,* 12**,** 1461-70.

CLOSE, D. W., DON PAUL, C., LANGAN, P. S., WILCE, M. C. J., TRAORE, D. A. K., HALFMANN, R., ROCHA, R. C., WALDO, G. S., PAYNE, R. J., RUCKER, J. B., PRESCOTT, M. & BRADBURY, A. R. M. 2014. TGP, an extremely stable, non-aggregating fluorescent protein created by structure-guided surface engineering. *Proteins: Structure, Function, and Bioinformatics***,** n/a-n/a.

COLLINS, K. D. 1997. Charge density-dependent strength of hydration and biological structure. *Biophys J,* 72**,** 65-76.

COLLINS, K. D. & WASHABAUGH, M. W. 1985. The Hofmeister effect and the behaviour of water at interfaces. *Q Rev Biophys,* 18**,** 323-422.

CONCHILLO-SOLE, O., DE GROOT, N., AVILES, F., VENDRELL, J., DAURA, X. & VENTURA, S. 2007. AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. *BMC Bioinformatics,* 8**,** 65.

CONSORTIUM, T. U. 2014. Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Research,* 42**,** D191-D198.

CORNELIS, P. 2000. Expressing genes in different Escherichia coli compartments. *Current Opinion in Biotechnology,* 11**,** 450-454.

COSTA, C. Z. F., DA ROSA, S. E. A. & DE CAMARGO, M. M. 2011. The Unfolded Protein Response: How Protein Folding Became a Restrictive Aspect for Innate Immunity and B Lymphocytes1. *Scandinavian Journal of Immunology,* 73**,** 436-448.

CUDNA, R. E. & DICKSON, A. J. 2003. Endoplasmic reticulum signaling as a determinant of recombinant protein expression. *Biotechnology and Bioengineering,* 81**,** 56-65.

CUERVO, A. M., WONG, E. S. P. & MARTINEZ-VICENTE, M. 2010. Protein degradation, aggregation, and misfolding. *Movement Disorders,* 25**,** S49-S54.

DALE, G. E., BROGER, C., LANGEN, H., ARCY, A. D. & STÜBER, D. 1994. Improving protein solubility through rationally designed amino acid replacements: solubilization of the trimethoprim-resistant type S1 dihydrofolate reductase. *Protein Engineering,* 7**,** 933-939.

DANTAS, G., KUHLMAN, B., CALLENDER, D., WONG, M. & BAKER, D. 2003. A Large Scale Test of Computational Protein Design: Folding and Stability of Nine Completely Redesigned Globular Proteins. *Journal of Molecular Biology,* 332**,** 449-460.

DAS, D. & GEORGIADIS, M. M. 2001. A directed approach to improving the solubility of Moloney murine leukemia virus reverse transcriptase. *Protein Sci,* 10**,** 1936-41.

DAUJOTYTĖ, D., VILKAITIS, G., MANELYTĖ, L., SKALICKY, J., SZYPERSKI, T. & KLIMAŠAUSKAS, S. 2003. Solubility engineering of the HhaI methyltransferase. *Protein Engineering,* 16**,** 295-301.

DAVIS, G. D., ELISEE, C., NEWHAM, D. M. & HARRISON, R. G. 1999. New fusion protein systems designed to give soluble expression in Escherichia coli. *Biotechnol Bioeng,* 65**,** 382-8.

DAVIS, J. M., ARAKAWA, T., STRICKLAND, T. W. & YPHANTIS, D. A. 1987. Characterization of recombinant human erythropoietin produced in Chinese hamster ovary cells. *Biochemistry,* 26**,** 2633-8.

DEBELJAK, N., FELDMAN, L., DAVIS, K. L., KOMEL, R. & SYTKOWSKI, A. J. 2006. Variability in the immunodetection of His-tagged recombinant proteins. *Anal Biochem,* 359**,** 216-23.

DELISA, M. P., TULLMAN, D. & GEORGIOU, G. 2003. Folding quality control in the export of proteins by the bacterial twin-arginine translocation pathway. *Proceedings of the National Academy of Sciences,* 100**,** 6115-6120.

DEPUYDT, M., LEONARD, S. E., VERTOMMEN, D., DENONCIN, K., MORSOMME, P., WAHNI, K., MESSENS, J., CARROLL, K. S. & COLLET, J.-F. 2009. A Periplasmic Reducing System Protects Single Cysteine Residues from Oxidation. *Science,* 326**,** 1109-1111.

DEREWENDA, Z. 2010. Application of protein engineering to enhance crystallizability and improve crystal properties. *Acta Crystallographica Section D,* 66**,** 604-615.

DEUERLING, E., PATZELT, H., VORDERWÜLBECKE, S., RAUCH, T., KRAMER, G., SCHAFFITZEL, E., MOGK, A., SCHULZE-SPECKING, A., LANGEN, H. & BUKAU, B. 2003. Trigger Factor and DnaK possess overlapping substrate pools and binding specificities. *Molecular Microbiology,* 47**,** 1317-1328.

DINGERMANN, T. 2008. Recombinant therapeutic proteins: Production platforms and challenges. *Biotechnology Journal,* 3**,** 90-97.

DOBSON, C. M., ŠALI, A. & KARPLUS, M. 1998. Protein Folding: A Perspective from Theory and Experiment. *Angewandte Chemie International Edition,* 37**,** 868-893.

DUNN, C. J. & GOA, K. L. 2001. Tenecteplase: a review of its pharmacology and therapeutic efficacy in patients with acute myocardial infarction. *Am J Cardiovasc Drugs,* 1**,** 51-66.

DYDA, F., HICKMAN, A., JENKINS, T., ENGELMAN, A., CRAIGIE, R. & DAVIES, D. 1994. Crystal structure of the catalytic domain of HIV-1 integrase: similarity to other polynucleotidyl transferases. *Science,* 266**,** 1981-1986.

DYSON, H. J., JENG, M.-F., TENNANT, L. L., SLABY, I., LINDELL, M., CUI, D.-S., KUPRIN, S. & HOLMGREN, A. 1997. Effects of Buried Charged Groups on Cysteine Thiol Ionization and Reactivity in Escherichia coli Thioredoxin: Structural and Functional Characterization of Mutants of Asp 26 and Lys 57†. *Biochemistry,* 36**,** 2622-2636.

EL-MAGHRABI, M. R., NOTO, F., WU, N. & MANES, N. 2001. 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase: suiting structure to need, in a family of tissue-specific enzymes. *Curr Opin Clin Nutr Metab Care,* 4**,** 411-8.

ELLGAARD, L. & FRICKEL, E.-M. 2003. Calnexin, calreticulin, and ERp57. *Cell Biochemistry and Biophysics,* 39**,** 223-247.

ELLGAARD, L. & HELENIUS, A. 2003. Quality control in the endoplasmic reticulum. *Nat Rev Mol Cell Biol,* 4**,** 181-191.

ELLIOTT, S. 2009. New molecules and formulations. *In:* ELLIOTT, S., FOOTE, M. & MOLINEUX, G. (eds.) *Erythropoietins, Erythropoietic Factors, and Erythropoiesis.* Birkhäuser Basel.

ELLIOTT, S., LORENZINI, T., ASHER, S., AOKI, K., BRANKOW, D., BUCK, L., BUSSE, L., CHANG, D., FULLER, J., GRANT, J., HERNDAY, N., HOKUM, M., HU, S., KNUDTEN, A., LEVIN, N., KOMOROWSKI, R., MARTIN, F., NAVARRO, R., OSSLUND, T., ROGERS, G., ROGERS, N., TRAIL, G. & EGRIE, J. 2003. Enhancement of therapeutic protein in vivo activities through glycoengineering. *Nat Biotechnol,* 21**,** 414-21.

ELLIOTT, S., LORENZINI, T., CHANG, D., BARZILAY, J. & DELORME, E. 1997. Mapping of the active site of recombinant human erythropoietin. *Blood,* 89**,** 493-502.

ESPOSITO, D. & CHATTERJEE, D. K. 2006. Enhancement of soluble protein expression through the use of fusion tags. *Current Opinion in Biotechnology,* 17**,** 353-358.

ETCHEGARAY, J.-P. & INOUYE, M. 1999. CspA, CspB, and CspG, Major Cold Shock Proteins ofEscherichia coli, Are Induced at Low Temperature under Conditions That Completely Block Protein Synthesis. *Journal of Bacteriology,* 181**,** 1827-1830.

FAN, D., LI, Q., KORANDO, L., JEROME, W. G. & WANG, J. 2004. A monomeric human apolipoprotein E carboxyl-terminal domain. *Biochemistry,* 43**,** 5055-64.

FANG, Y. & FANG, J. 2013. Discrimination of soluble and aggregation-prone proteins based on sequence information. *Mol Biosyst,* 9**,** 806-11.

FARADY, C. J., SELLERS, B. D., JACOBSON, M. P. & CRAIK, C. S. 2009. Improving the species cross-reactivity of an antibody using computational design. *Bioorg Med Chem Lett,* 19**,** 3744-7.

FAREWELL, A. & NEIDHARDT, F. C. 1998. Effect of temperature on in vivo protein synthetic capacity in Escherichia coli. *J Bacteriol,* 180**,** 4704-10.

FERNANDEZ-ESCAMILLA, A.-M., ROUSSEAU, F., SCHYMKOWITZ, J. & SERRANO, L. 2004. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat Biotech,* 22**,** 1302-1306.

FERRER, M., CHERNIKOVA, T. N., YAKIMOV, M. M., GOLYSHIN, P. N. & TIMMIS, K. N. 2003. Chaperonins govern growth of Escherichia coli at low temperatures. *Nat Biotechnol,* 21**,** 1266-7.

FINK, A. L. 1998. Protein aggregation: folding aggregates, inclusion bodies and amyloid. *Fold Des,* 3**,** R9-23.

FISCHER, B., SUMNER, I. & GOODENOUGH, P. 1993. Isolation, renaturation, and formation of disulfide bonds of eukaryotic proteins expressed in Escherichia coli as inclusion bodies. *Biotechnology and Bioengineering,* 41**,** 3-13.

FOTIOU, F., ARAVIND, S., WANG, P. P. & NERAPUSEE, O. 2009. Impact of illegal trade on the quality of epoetin alfa in Thailand. *Clin Ther,* 31**,** 336-46.

FOWLER, S. B., POON, S., MUFF, R., CHITI, F., DOBSON, C. M. & ZURDO, J. 2005. Rational design of aggregation-resistant bioactive peptides: reengineering human calcitonin. *Proc Natl Acad Sci U S A,* 102**,** 10105-10.

FRADET-TURCOTTE, A., CANNY, M. D., ESCRIBANO-DIAZ, C., ORTHWEIN, A., LEUNG, C. C., HUANG, H., LANDRY, M. C., KITEVSKI-LEBLANC, J., NOORDERMEER, S. M., SICHERI, F. & DUROCHER, D. 2013. 53BP1 is a reader of the DNA-damage-induced H2A Lys 15 ubiquitin mark. *Nature,* 499**,** 50-4.

FRAND, A. R., CUOZZO, J. W. & KAISER, C. A. 2000. Pathways for protein disulphide bond formation. *Trends in Cell Biology,* 10**,** 203-210.

FREEDMAN, R. B. 1989. Protein disulfide isomerase: Multiple roles in the modification of nascent secretory proteins. *Cell,* 57**,** 1069-1072.

FRICKEL, E.-M., FREI, P., BOUVIER, M., STAFFORD, W. F., HELENIUS, A., GLOCKSHUBER, R. & ELLGAARD, L. 2004. ERp57 Is a Multifunctional Thiol-Disulfide Oxidoreductase. *Journal of Biological Chemistry,* 279**,** 18277-18287.

FRYDMAN, J. 2001. FOLDING OF NEWLY TRANSLATED PROTEINS IN VIVO: The Role of Molecular Chaperones. *Annual Review of Biochemistry,* 70**,** 603-647.

FRYDMAN, J., NIMMESGERN, E., OHTSUKA, K. & HARTL, F. U. 1994. Folding of nascent polypeptide chains in a high molecular mass assembly with molecular chaperones. *Nature,* 370**,** 111-117.

GALLAGHER, D. T., SMITH, N. N., KIM, S.-K., ROBINSON, H. & REDDY, P. T. 2009. Protein Crystal Engineering of YpAC-IV using the Strategy of Excess Charge Reduction. *Crystal growth & design,* 9**,** 3570-3574.

GASTEIGER, E., HOOGLAND, C., GATTIKER, A., DUVAUD, S. E., WILKINS, M., APPEL, R. & BAIROCH, A. 2005. Protein Identification and Analysis Tools on the ExPASy Server. *In:* WALKER, J. (ed.) *The Proteomics Protocols Handbook.* Humana Press.

GATTI, M., PINATO, S., MASPERO, E., SOFFIENTINI, P., POLO, S. & PENENGO, L. 2012. A novel ubiquitin mark at the N-terminal tail of histone H2As targeted by RNF168 ubiquitin ligase. *Cell Cycle,* 11**,** 2538-44.

GOLDBERG, D. S., BISHOP, S. M., SHAH, A. U. & SATHISH, H. A. 2011. Formulation development of therapeutic monoclonal antibodies using high-throughput fluorescence and static light scattering techniques: Role of conformational and colloidal stability. *Journal of Pharmaceutical Sciences,* 100**,** 1306-1315.

GOPAL, G. & KUMAR, A. 2013. Strategies for the Production of Recombinant Protein in Escherichia coli. *The Protein Journal,* 32**,** 419-425.

GRAUMANN, P. & MARAHIEL, M. A. 1997. Effects of heterologous expression of CspB, the major cold shock protein of Bacillus subtillis, on protein synthesis in Escherichia coli. *Mol Gen Genet,* 253**,** 745-52.

GREAVES, R. B. & WARWICKER, J. 2007. Mechanisms for stabilisation and the maintenance of solubility in proteins from thermophiles. *BMC Struct Biol,* 7**,** 18.

GUEX, N. & PEITSCH, M. C. 1997. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis,* 18**,** 2714-23.

GUSTAFSSON, C., GOVINDARAJAN, S. & MINSHULL, J. 2004. Codon bias and heterologous protein expression. *Trends in Biotechnology,* 22**,** 346-353.

HAMRANG, Z., RATTRAY, N. J. W. & PLUEN, A. 2013. Proteins behaving badly: emerging technologies in profiling biopharmaceutical aggregation. *Trends in Biotechnology,* 31**,** 448-458.

HANSEN, W. J., COWAN, N. J. & WELCH, W. J. 1999. Prefoldin–Nascent Chain Complexes in the Folding of Cytoskeletal Proteins. *The Journal of Cell Biology,* 145**,** 265-277.

HANSMANN, U. H. 2008. Toward reliable simulations of protein folding, misfolding and aggregation. *Prog Mol Biol Transl Sci,* 84**,** 39-55.

HARTL, F. U. 1996. Molecular chaperones in cellular protein folding. *Nature,* 381**,** 571-580.

HASEMANN, C. A., ISTVAN, E. S., UYEDA, K. & DEISENHOFER, J. 1996. The crystal structure of the bifunctional enzyme 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase reveals distinct domain homologies. *Structure,* 4**,** 1017-29.

HATFIELD, G. W. & ROTH, D. A. 2007. Optimizing scaleup yield for protein production: Computationally Optimized DNA Assembly (CODA) and Translation Engineering™. *In:* EL-GEWELY, M. R. (ed.) *Biotechnology Annual Review.* Elsevier.

HESTERKAMP, T., HAUSER, S., LÜTCKE, H. & BUKAU, B. 1996. Escherichia coli trigger factor is a prolyl isomerase that associates with nascent polypeptide chains. *Proceedings of the National Academy of Sciences,* 93**,** 4437-4441.

HIGUCHI, M., OH-EDA, M., KUBONIWA, H., TOMONOH, K., SHIMONAKA, Y. & OCHI, N. 1992. Role of sugar chains in the expression of the biological activity of human erythropoietin. *J Biol Chem,* 267**,** 7703-9.

HIROSE, S., KAWAMURA, Y., YOKOTA, K., KUROITA, T., NATSUME, T., KOMIYA, K., TSUTSUMI, T., SUWA, Y., ISOGAI, T., GOSHIMA, N. & NOGUCHI, T. 2011. Statistical analysis of features associated with protein expression/solubility in an in vivo Escherichia coli expression system and a wheat germ cell-free expression system. *Journal of Biochemistry,* 150**,** 73-81.

HIROSE, S. & NOGUCHI, T. 2013. ESPRESSO: a system for estimating protein expression and solubility in protein expression systems. *Proteomics,* 13**,** 1444-56.

HIRSCH, I. B. 2005. Insulin Analogues. *New England Journal of Medicine,* 352**,** 174-183.

HOFFMANN, F. & RINAS, U. 2004. Roles of Heat-Shock Chaperones in the Production of Recombinant Proteins in Escherichia coli. *Physiological Stress Responses in Bioprocesses.* Springer Berlin Heidelberg.

HORWICH, A. L., LOW, K. B., FENTON, W. A., HIRSHFIELD, I. N. & FURTAK, K. 1993. Folding in vivo of bacterial cytoplasmic proteins: Role of GroEL. *Cell,* 74**,** 909-917.

HUANG, H. L., CHAROENKWAN, P., KAO, T. F., LEE, H. C., CHANG, F. L., HUANG, W. L., HO, S. J., SHU, L. S., CHEN, W. L. & HO, S. Y. 2012. Prediction and analysis of protein solubility using a novel scoring card method with dipeptide composition. *BMC Bioinformatics,* 13 Suppl 17**,** S3.

HUE, L. & ROUSSEAU, G. G. 1993. Fructose 2,6-bisphosphate and the control of glycolysis by growth factors, tumor promoters and oncogenes. *Advances in Enzyme Regulation,* 33**,** 97-110.

HUNG, S. C., KANG, M.-S. & KIEFF, E. 2001. Maintenance of Epstein–Barr virus (EBV) oriP-based episomes requires EBV-encoded nuclear antigen-1 chromosome-binding domains, which can be replaced by high-mobility group-I or histone H1. *Proceedings of the National Academy of Sciences,* 98**,** 1865-1870.

HWANG, I. & PARK, S. 2008. Computational design of protein therapeutics. *Drug Discovery Today: Technologies,* 5**,** e43-e48.

IDICULA-THOMAS, S. & BALAJI, P. V. 2005. Understanding the relationship between the primary structure of proteins and its propensity to be soluble on overexpression in Escherichia coli. *Protein Sci,* 14**,** 582-92.

JÄCKEL, C., KAST, P. & HILVERT, D. 2008. Protein Design by Directed Evolution. *Annual Review of Biophysics,* 37**,** 153-173.

JACOBS, K., SHOEMAKER, C., RUDERSDORF, R., NEILL, S. D., KAUFMAN, R. J., MUFSON, A., SEEHRA, J., JONES, S. S., HEWICK, R., FRITSCH, E. F. & ET AL. 1985. Isolation and characterization of genomic and cDNA clones of human erythropoietin. *Nature,* 313**,** 806-10.

JELKMANN, W. 2013. Physiology and pharmacology of erythropoietin. *Transfus Med Hemother,* 40**,** 302-9.

JENKINS, N. 2007. Modifications of therapeutic proteins: challenges and prospects. *Cytotechnology,* 53**,** 121-125.

JENKINS, N., MELEADY, P., TYTHER, R. & MURPHY, L. 2009. Strategies for analysing and improving the expression and quality of recombinant proteins made in mammalian cells. *Biotechnology and Applied Biochemistry,* 53**,** 73-83.

JENKINS, N., MURPHY, L. & TYTHER, R. 2008. Post-translational Modifications of Recombinant Proteins: Significance for Biopharmaceuticals. *Molecular Biotechnology,* 39**,** 113-118.

JENKINS, T. M., HICKMAN, A. B., DYDA, F., GHIRLANDO, R., DAVIES, D. R. & CRAIGIE, R. 1995. Catalytic domain of human immunodeficiency virus type 1

integrase: identification of a soluble mutant by systematic replacement of hydrophobic residues. *Proc Natl Acad Sci U S A,* 92**,** 6057-61.

JEONG, T. H., SON, Y. J., RYU, H. B., KOO, B. K., JEONG, S. M., HOANG, P., DO, B. H., SONG, J. A., CHONG, S. H., ROBINSON, R. C. & CHOE, H. 2014. Soluble expression and partial purification of recombinant human erythropoietin from E. coli. *Protein Expr Purif,* 95**,** 211-8.

JEONG, Y. T., CHOI, O., LIM, H. R., SON, Y. D., KIM, H. J. & KIM, J. H. 2008. Enhanced sialylation of recombinant erythropoietin in CHO cells by human glycosyltransferase expression. *J Microbiol Biotechnol,* 18**,** 1945-52.

JONASSON, P., LILJEQVIST, S., NYGREN, P.-A. K. & STÅHL, S. 2002. Genetic design for facilitated production and recovery of recombinant proteins in Escherichia coli. *Biotechnology and Applied Biochemistry,* 35**,** 91-105.

KAMIONER, D. 2012. Erythropoietin biosimilars currently available in hematology-oncology. *Target Oncol,* 7 Suppl 1**,** S25-8.

KANE, J. F. & HARTLEY, D. L. 1988. Formation of recombinant protein inclusion bodies in Escherichia coli. *Trends in Biotechnology,* 6**,** 95-101.

KATTI, S. K., LEMASTER, D. M. & EKLUND, H. 1990. Crystal structure of thioredoxin from Escherichia coli at 1.68 Å resolution. *Journal of Molecular Biology,* 212**,** 167-184.

KHAN, M. A., ISLAM, M. M. & KURODA, Y. 2013. Analysis of protein aggregation kinetics using short amino acid peptide tags. *Biochim Biophys Acta,* 1834**,** 2107-15.

KIM, S. G., CAVALIER, M., EL-MAGHRABI, M. R. & LEE, Y. H. 2007. A direct substrate-substrate interaction found in the kinase domain of the bifunctional enzyme, 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase. *J Mol Biol,* 370**,** 14-26.

KIM, S. G., MANES, N. P., EL-MAGHRABI, M. R. & LEE, Y. H. 2006. Crystal structure of the hypoxia-inducible form of 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase (PFKFB3): a possible new target for cancer therapy. *J Biol Chem,* 281**,** 2939-44.

KIM, Y. E., HIPP, M. S., BRACHER, A., HAYER-HARTL, M. & ULRICH HARTL, F. 2013. Molecular Chaperone Functions in Protein Folding and Proteostasis. *Annual Review of Biochemistry,* 82**,** 323-355.

KITAGAWA, M., ARA, T., ARIFUZZAMAN, M., IOKA-NAKAMICHI, T., INAMOTO, E., TOYONAGA, H. & MORI, H. 2006. Complete set of ORF clones of Escherichia coli ASKA library (A Complete Set of E. coli K-12 ORF Archive): Unique Resources for Biological Research. *DNA Research,* 12**,** 291-299.

KITAMURA, T., TANGE, T., TERASAWA, T., CHIBA, S., KUWAKI, T., MIYAGAWA, K., PIAO, Y.-F., MIYAZONO, K., URABE, A. & TAKAKU, F. 1989. Establishment and characterization of a unique human cell line that proliferates dependently on GM-CSF, IL-3, or erythropoietin. *Journal of Cellular Physiology,* 140**,** 323-334.

KLUS, P., BOLOGNESI, B., AGOSTINI, F., MARCHESE, D., ZANZONI, A. & TARTAGLIA, G. G. 2014. The cleverSuite approach for protein characterization: predictions of structural properties, solubility, chaperone requirements and RNA-binding abilities. *Bioinformatics,* 30**,** 1601-8.

KOBATA, A. 1992. Structures and functions of the sugar chains of glycoproteins. *European Journal of Biochemistry,* 209**,** 483-501.

KRAMER, R. M., SHENDE, V. R., MOTL, N., PACE, C. N. & SCHOLTZ, J. M. 2012. Toward a molecular understanding of protein solubility: increased negative surface charge correlates with increased solubility. *Biophys J,* 102**,** 1907-15.

KRANTZ, S. B. 1991. Erythropoietin. *Blood,* 77**,** 419-34.

KUNTZ, I. D. 1971. Hydration of macromolecules. III. Hydration of polypeptides. *Journal of the American Chemical Society,* 93**,** 514-516.

KWAKS, T. H. J. & OTTE, A. P. 2006. Employing epigenetics to augment the expression of therapeutic proteins in mammalian cells. *Trends in Biotechnology,* 24**,** 137-142.

LAI, P. H., EVERETT, R., WANG, F. F., ARAKAWA, T. & GOLDWASSER, E. 1986. Structural characterization of human erythropoietin. *J Biol Chem,* 261**,** 3116-21.

LAWSON, A. J., WALKER, E. A., WHITE, S. A., DAFFORN, T. R., STEWART, P. M. & RIDE, J. P. 2009. Mutations of key hydrophobic surface residues of 11 beta-hydroxysteroid dehydrogenase type 1 increase solubility and monodispersity in a bacterial expression system. *Protein Sci,* 18**,** 1552-63.

LE FOURN, V., GIROD, P.-A., BUCETA, M., REGAMEY, A. & MERMOD, N. 2014. CHO cell engineering to prevent polypeptide aggregation and improve therapeutic protein secretion. *Metabolic Engineering,* 21**,** 91-102.

LEE, P. A., TULLMAN-ERCEK, D. & GEORGIOU, G. 2006. The Bacterial Twin-Arginine Translocation Pathway. *Annual Review of Microbiology,* 60**,** 373-395.

LI, Y., YAN, Y., ZUGAY-MURPHY, J., XU, B., COLE, J. L., WITMER, M., FELOCK, P., WOLFE, A., HAZUDA, D., SARDANA, M. K., CHEN, Z., KUO, L. C. & SARDANA, V. V. 1999. Purification, solution properties and crystallization of SIV integrase containing a continuous core and C-terminal domain. *Acta Crystallogr D Biol Crystallogr,* 55**,** 1906-10.

LIN, K., KURLAND, I., XU, L. Z., LANGE, A. J., PILKIS, J., EL-MAGHRABI, M. R. & PILKIS, S. J. 1990. Expression of mammalian liver glycolytic/gluconeogenic enzymes in Escherichia coli: recovery of active enzyme is strain and temperature dependent. *Protein Expr Purif,* 1**,** 169-76.

LIPPOW, S. M., WITTRUP, K. D. & TIDOR, B. 2007. Computational design of antibody-affinity improvement beyond in vivo maturation. *Nat Biotechnol,* 25**,** 1171-6.

LIRAS, A. 2008. Recombinant proteins in therapeutics: haemophilia treatment as an example. *International Archives of Medicine,* 1**,** 4-4.

LISOWSKA, E. 2002. The role of glycosylation in protein antigenic properties. *Cell Mol Life Sci,* 59**,** 445-55.

LOBSTEIN, J., EMRICH, C., JEANS, C., FAULKNER, M., RIGGS, P. & BERKMEN, M. 2012. SHuffle, a novel Escherichia coli protein expression strain capable of correctly folding disulfide bonded proteins in its cytoplasm. *Microbial Cell Factories,* 11**,** 56.

MAGNAN, C. N., RANDALL, A. & BALDI, P. 2009. SOLpro: accurate sequence-based prediction of protein solubility. *Bioinformatics,* 25**,** 2200-7.

MAHLER, H.-C., FRIESS, W., GRAUSCHOPF, U. & KIESE, S. 2009. Protein aggregation: Pathways, induction factors and analysis. *Journal of Pharmaceutical Sciences,* 98**,** 2909-2934.

MAHMOUD, K. 2007. Recombinant protein production: strategic technology and a vital research tool. *Res. J. Cell Mol. Biol.,* 1**,** 9-22.

MAJHI, P. R., GANTA, R. R., VANAM, R. P., SEYREK, E., GIGER, K. & DUBIN, P. L. 2006. Electrostatically driven protein aggregation: beta-lactoglobulin at low ionic strength. *Langmuir,* 22**,** 9150-9.

MALAKAUSKAS, S. M. & MAYO, S. L. 1998. Design, structure and stability of a hyperthermophilic protein variant. *Nat Struct Mol Biol,* 5**,** 470-475.

MARKOSSIAN, K. A. & KURGANOV, B. I. 2004. Protein Folding, Misfolding, and Aggregation. Formation of Inclusion Bodies and Aggresomes. *Biochemistry (Moscow),* 69**,** 971-984.

MARSTON, F. A. 1986. The purification of eukaryotic polypeptides synthesized in Escherichia coli. *Biochem J,* 240**,** 1-12.

MATASCI, M., HACKER, D. L., BALDI, L. & WURM, F. M. 2008. Recombinant therapeutic protein production in cultivated mammalian cells: current status and future prospects. *Drug Discovery Today: Technologies,* 5**,** e37-e42.

MATTIROLI, F. & SIXMA, T. K. 2014. Lysine-targeting specificity in ubiquitin and ubiquitin-like modification pathways. *Nat Struct Mol Biol,* 21**,** 308-16.

MATTIROLI, F., UCKELMANN, M., SAHTOE, D. D., VAN DIJK, W. J. & SIXMA, T. K. 2014. The nucleosome acidic patch plays a critical role in RNF168-dependent ubiquitination of histone H2A. *Nat Commun,* 5.

MAYER, M. P. 2010. Gymnastics of Molecular Chaperones. *Molecular Cell,* 39**,** 321-331.

MCKOY, J. M., STONECASH, R. E., COURNOYER, D., ROSSERT, J., NISSENSON, A. R., RAISCH, D. W., CASADEVALL, N. & BENNETT, C. L. 2008. Epoetin-associated pure red cell aplasia: past, present, and future considerations. *Transfusion,* 48**,** 1754-62.

MEADOW, N. D., FOX, D. K. & ROSEMAN, S. 1990. The Bacterial Phosphoenol-Pyruvate: Glycose Phosphotransferase System. *Annual Review of Biochemistry,* 59**,** 497-542.

MEISSNER, P., PICK, H., KULANGARA, A., CHATELLARD, P., FRIEDRICH, K. & WURM, F. M. 2001. Transient gene expression: Recombinant protein production with suspension-adapted HEK293-EBNA cells. *Biotechnology and Bioengineering,* 75**,** 197-203.

MIKHAIL, A. & FAROUK, M. 2013. Epoetin biosimilars in Europe: five years on. *Adv Ther,* 30**,** 28-40.

MITRAKI, A. & KING, J. 1989. Protein Folding Intermediates and Inclusion Body Formation. *Nat Biotech,* 7**,** 690-697.

MOGK, A., MAYER, M. P. & DEUERLING, E. 2002. Mechanisms of Protein Folding: Molecular Chaperones and Their Application in Biotechnology. *ChemBioChem,* 3**,** 807-814.

MOHAN, C., PARK, S. H., CHUNG, J. Y. & LEE, G. M. 2007. Effect of doxycycline-regulated protein disulfide isomerase expression on the specific productivity of recombinant CHO cells: Thrombopoietin and antibody. *Biotechnology and Bioengineering,* 98**,** 611-615.

MOSAVI, L. K. & PENG, Z. Y. 2003. Structure-based substitutions for increased solubility of a designed protein. *Protein Engineering,* 16**,** 739-745.

NARHI, L. O., ARAKAWA, T., AOKI, K., WEN, J., ELLIOTT, S., BOONE, T. & CHEETHAM, J. 2001. Asn to Lys mutations at three sites which are N-glycosylated in the mammalian protein decrease the aggregation of Escherichia coli-derived erythropoietin. *Protein Eng,* 14**,** 135-40.

NARHI, L. O., ARAKAWA, T., AOKI, K. H., ELMORE, R., ROHDE, M. F., BOONE, T. & STRICKLAND, T. W. 1991. The effect of carbohydrate on the structure and stability of erythropoietin. *J Biol Chem,* 266**,** 23022-6.

NASREEN, A., VOGT, M., KIM, H. J., EICHINGER, A. & SKERRA, A. 2006. Solubility engineering and crystallization of human apolipoprotein D. *Protein Sci,* 15**,** 190-9.

NEWKIRK, K., FENG, W., JIANG, W., TEJERO, R., EMERSON, S. D., INOUYE, M. & MONTELIONE, G. T. 1994. Solution NMR structure of the major cold shock protein (CspA) from Escherichia coli: identification of a binding epitope for DNA. *Proceedings of the National Academy of Sciences of the United States of America,* 91**,** 5114-5118.

NIU, X. H., HU, X. H., SHI, F. & XIA, J. B. 2012. Predicting protein solubility by the general form of Chou's pseudo amino acid composition: approached from chaos game representation and fractal dimension. *Protein Pept Lett,* 19**,** 940-8.

NIWA, T., KANAMORI, T., UEDA, T. & TAGUCHI, H. 2012. Global analysis of chaperone effects using a reconstituted cell-free translation system. *Proceedings of the National Academy of Sciences,* 109**,** 8937-8942.

NIWA, T., YING, B. W., SAITO, K., JIN, W., TAKADA, S., UEDA, T. & TAGUCHI, H. 2009. Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of Escherichia coli proteins. *Proc Natl Acad Sci U S A,* 106**,** 4201-6.

NYATHI, Y., WILKINSON, B. M. & POOL, M. R. 2013. Co-translational targeting and translocation of proteins to the endoplasmic reticulum. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research,* 1833**,** 2392-2402.

PALADE, G. 1975. Intracellular Aspects of the Process of Protein Synthesis. *Science,* 189**,** 867.

PALOMARES, L., ESTRADA-MONCADA, S. & RAMÍREZ, O. 2004. Production of Recombinant Proteins. *In:* BALBÁS, P. & LORENCE, A. (eds.) *Recombinant Gene Expression.* Humana Press.

PARHAM, J., KOST, T. & HUTCHINS, J. 2001. Effects of pCIneo and pCEP4 expression vectors on transient and stable protein production in human and simian cell lines. *Cytotechnology,* 35**,** 181-187.

PARK, S., YANG, X. & SAVEN, J. G. 2004. Advances in computational protein design. *Curr Opin Struct Biol,* 14**,** 487-94.

PARK, S. S., PARK, J., KO, J., CHEN, L., MERIAGE, D., CROUSE-ZEINEDDINI, J., WONG, W. & KERWIN, B. A. 2009. Biochemical assessment of erythropoietin products from Asia versus US Epoetin alfa manufactured by Amgen. *Journal of Pharmaceutical Sciences,* 98**,** 1688-1699.

PAROUTIS, P., TOURET, N. & GRINSTEIN, S. 2004. The pH of the secretory pathway: measurement, determinants, and regulation. *Physiology (Bethesda),* 19**,** 207-15.

PATEL, S. B., CAMERON, P. M., FRANTZ-WATTLEY, B., O'NEILL, E., BECKER, J. W. & SCAPIN, G. 2004. Lattice stabilization and enhanced diffraction in human p38 alpha crystals by protein engineering. *Biochim Biophys Acta,* 1696**,** 67-73.

PAVLOU, A. K. & REICHERT, J. M. 2004. Recombinant protein therapeutics[mdash]success rates, market trends and values to 2010. *Nat Biotech,* 22**,** 1513-1519.

PENG, R.-W. & FUSSENEGGER, M. 2009. Engineering the Secretory Pathway in Mammalian Cells. *In:* AL-RUBEAI, M. (ed.) *Cell Line Development.* Springer Netherlands.

PEPINSKY, R. B., SILVIAN, L., BERKOWITZ, S. A., FARRINGTON, G., LUGOVSKOY, A., WALUS, L., ELDREDGE, J., CAPILI, A., MI, S., GRAFF, C. & GARBER, E. 2010. Improving the solubility of anti-LINGO-1 monoclonal antibody Li33 by isotype switching and targeted mutagenesis. *Protein Sci,* 19**,** 954-66.

PERCHIACCA, J. M., LADIWALA, A. R., BHATTACHARYA, M. & TESSIER, P. M. 2012. Aggregation-resistant domain antibodies engineered with charged mutations near the edges of the complementarity-determining regions. *Protein Eng Des Sel,* 25**,** 591-601.

PETI, W. & PAGE, R. 2007. Strategies to maximize heterologous protein expression in Escherichia coli with minimal cost. *Protein Expr Purif,* 51**,** 1-10.

PILKIS, S. J., REGEN, D. M., STEWART, H. B., PILKIS, J., PATE, T. M. & EL-MAGHRABI, M. R. 1984. Evidence for two catalytic sites on 6-phosphofructo-2-kinase/fructose 2,6-bisphosphatase. Dynamics of substrate exchange and phosphoryl enzyme formation. *J Biol Chem,* 259**,** 949-58.

PRADITPORNSILPA, K., TIRANATHANAGUL, K., KUPATAWINTU, P., JOOTAR, S., INTRAGUMTORNCHAI, T., TUNGSANGA, K., TEERAPORNLERTRATT, T., LUMLERTKUL, D., TOWNAMCHAI, N., SUSANTITAPHONG, P., KATAVETIN, P., KANJANABUCH, T., AVIHINGSANON, Y. & EIAM-ONG, S. 2011. Biosimilar

recombinant human erythropoietin induces the production of neutralizing antibodies. *Kidney Int,* 80**,** 88-92.

PRICE, W. N., 2ND, HANDELMAN, S. K., EVERETT, J. K., TONG, S. N., BRACIC, A., LUFF, J. D., NAUMOV, V., ACTON, T., MANOR, P., XIAO, R., ROST, B., MONTELIONE, G. T. & HUNT, J. F. 2011. Large-scale experimental studies show unexpected amino acid effects on protein expression and solubility in vivo in E. coli. *Microb Inform Exp,* 1**,** 6.

PRILUSKY, J., FELDER, C. E., ZEEV-BEN-MORDEHAI, T., RYDBERG, E. H., MAN, O., BECKMANN, J. S., SILMAN, I. & SUSSMAN, J. L. 2005. FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics,* 21**,** 3435-8.

RACE, P. R., SOLOVYOVA, A. S. & BANFIELD, M. J. 2007. Conformation of the EPEC Tir protein in solution: investigating the impact of serine phosphorylation at positions 434/463. *Biophys J,* 93**,** 586-96.

RATANJI, K. D., DERRICK, J. P., DEARMAN, R. J. & KIMBER, I. 2014. Immunogenicity of therapeutic proteins: influence of aggregation. *J Immunotoxicol,* 11**,** 99-109.

RECNY, M. A., SCOBLE, H. A. & KIM, Y. 1987. Structural characterization of natural human urinary and recombinant DNA-derived erythropoietin. Identification of des-arginine 166 erythropoietin. *J Biol Chem,* 262**,** 17156-63.

ROBINSON, C. & BOLHUIS, A. 2001. Protein targeting by the twin-arginine translocation pathway. *Nat Rev Mol Cell Biol,* 2**,** 350-356.

ROOSILD, T. P. & CHOE, S. 2005. Redesigning an integral membrane K+ channel into a soluble protein. *Protein Engineering Design and Selection,* 18**,** 79-84.

ROOSILD, T. P., VEGA, M., CASTRONOVO, S. & CHOE, S. 2006. Characterization of the family of Mistic homologues. *BMC Struct Biol,* 6**,** 10.

SACKMANN, E. 1995. Chapter 1 Biological membranes architecture and function. *In:* LIPOWSKY, R. & SACKMANN, E. (eds.) *Handbook of Biological Physics.* North-Holland.

SAHDEV, S., KHATTAR, S. & SAINI, K. 2008. Production of active eukaryotic proteins through bacterial expression systems: a review of the existing biotechnology strategies. *Molecular and Cellular Biochemistry,* 307**,** 249-264.

SAIBIL, H. 2013. Chaperone machines for protein folding, unfolding and disaggregation. *Nat Rev Mol Cell Biol,* 14**,** 630-642.

SAKAKIBARA, R., KATO, M., OKAMURA, N., NAKAGAWA, T., KOMADA, Y., TOMINAGA, N., SHIMOJO, M. & FUKASAWA, M. 1997. Characterization of a human placental fructose-6-phosphate, 2-kinase/fructose-2,6-bisphosphatase. *J Biochem,* 122**,** 122-8.

SAMAK, T., GUNTER, D. & WANG, Z. 2012. Prediction of protein solubility in E. coli. *Proceedings of the 2012 IEEE 8th International Conference on E-Science (e-Science).* IEEE Computer Society.

SAMALI, A., FITZGERALD, U., DEEGAN, S. & GUPTA, S. 2010. Methods for Monitoring Endoplasmic Reticulum Stress and the Unfolded Protein Response. *International Journal of Cell Biology,* 2010**,** 11.

SCHEIN, C. H. & NOTEBORN, M. H. M. 1988. Formation of Soluble Recombinant Proteins in Escherichia Coli is Favored by Lower Growth Temperature. *Nat Biotech,* 6**,** 291-294.

SCHINDELIN, H., JIANG, W., INOUYE, M. & HEINEMANN, U. 1994. Crystal structure of CspA, the major cold shock protein of Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America,* 91**,** 5119-5123.

SCHINDELIN, H., MARAHIEL, M. A. & HEINEMANN, U. 1993. Universal nucleic acid-binding domain revealed by crystal structure of the B. subtilis major cold-shock protein. *Nature,* 364**,** 164-168.

SCHNUCHEL, A., WILTSCHECK, R., CZISCH, M., HERRLER, M., WLLLLMSKY, G., GRAUMANN, P., MARAHIEL, M. A. & HOLAK, T. A. 1993. Structure in solution of the major cold-shock protein from Bacillus subtilis. *Nature,* 364**,** 169-171.

SCHRÖDER, M. 2008. Engineering eukaryotic protein factories. *Biotechnology Letters,* 30**,** 187-196.

SCHRÖDER, M. & KAUFMAN, R. J. 2005. THE MAMMALIAN UNFOLDED PROTEIN RESPONSE. *Annual Review of Biochemistry,* 74**,** 739-789.

SCHRÖDINGER, L. L. C. 2010. The PyMOL Molecular Graphics System, Version 1.3r1.

SCHUMANN, W. & FERREIRA, L. C. S. 2004. Production of recombinant proteins in Escherichia coli. *Genetics and Molecular Biology,* 27**,** 442-453.

SCHYMKOWITZ, J., BORG, J., STRICHER, F., NYS, R., ROUSSEAU, F. & SERRANO, L. 2005. The FoldX web server: an online force field. *Nucleic Acids Res,* 33**,** W382-8.

SEIDL, A., HAINZL, O., RICHTER, M., FISCHER, R., BOHM, S., DEUTEL, B., HARTINGER, M., WINDISCH, J., CASADEVALL, N., LONDON, G. M. & MACDOUGALL, I. 2012. Tungsten-induced denaturation and aggregation of epoetin alfa during primary packaging as a cause of immunogenicity. *Pharm Res,* 29**,** 1454-67.

SEO, M., KIM, J. D., NEAU, D., SEHGAL, I. & LEE, Y. H. 2011. Structure-based development of small molecule PFKFB3 inhibitors: a framework for potential cancer therapeutic agents targeting the Warburg effect. *PLoS One,* 6**,** e24179.

SEVASTSYANOVICH, Y., ALFASI, S., OVERTON, T., HALL, R., JONES, J., HEWITT, C. & COLE, J. 2009. Exploitation of GFP fusion proteins and stress avoidance as a generic strategy for the production of high-quality recombinant proteins. *FEMS Microbiol Lett,* 299**,** 86-94.

SHELDON, B., RUSSELL-JONES, D. & WRIGHT, J. 2009. Insulin analogues: an example of applied medical science. *Diabetes Obes Metab,* 11**,** 5-19.

SHIMIZU, Y., KANAMORI, T. & UEDA, T. 2005. Protein synthesis by pure translation systems. *Methods,* 36**,** 299-304.

SKIBELI, V., NISSEN-LIE, G. & TORJESEN, P. 2001. *Sugar profiling proves that human serum erythropoietin differs from recombinant human erythropoietin*.

SLOVIC, A. M., KONO, H., LEAR, J. D., SAVEN, J. G. & DEGRADO, W. F. 2004. Computational design of water-soluble analogues of the potassium channel KcsA. *Proceedings of the National Academy of Sciences of the United States of America,* 101**,** 1828-1833.

SLOVIC, A. M., SUMMA, C. M., LEAR, J. D. & DEGRADO, W. F. 2003. Computational design of a water-soluble analog of phospholamban. *Protein Sci,* 12**,** 337-48.

SMIALOWSKI, P., DOOSE, G., TORKLER, P., KAUFMANN, S. & FRISHMAN, D. 2012. PROSO II--a new method for protein solubility prediction. *FEBS J,* 279**,** 2192-200.

SMIALOWSKI, P., MARTIN-GALIANO, A. J., MIKOLAJKA, A., GIRSCHICK, T., HOLAK, T. A. & FRISHMAN, D. 2007. Protein solubility: sequence based prediction and experimental verification. *Bioinformatics,* 23**,** 2536-42.

SOCKOLOSKY, J. T. & SZOKA, F. C. 2013. Periplasmic production via the pET expression system of soluble, bioactive human growth hormone. *Protein Expression and Purification,* 87**,** 129-135.

SOKALINGAM, S., RAGHUNATHAN, G., SOUNDRARAJAN, N. & LEE, S.-G. 2012. A Study on the Effect of Surface Lysine to Arginine Mutagenesis on Protein Stability and Structure Using Green Fluorescent Protein. *PLoS ONE,* 7**,** e40410.

SOLÁ, R. & GRIEBENOW, K. 2010. Glycosylation of Therapeutic Proteins. *BioDrugs,* 24**,** 9-21.
SOLÁ, R. J. & GRIEBENOW, K. 2009. Effects of glycosylation on the stability of protein pharmaceuticals. *Journal of Pharmaceutical Sciences,* 98**,** 1223-1245.
SØRENSEN, H. P. & MORTENSEN, K. K. 2005a. Advanced genetic strategies for recombinant protein expression in Escherichia coli. *Journal of Biotechnology,* 115**,** 113-128.
SØRENSEN, H. P. & MORTENSEN, K. K. 2005b. Soluble expression of recombinant proteins in the cytoplasm of Escherichia coli. *Microbial Cell Factories,* 4**,** 1-1.
SPECHT, E., MIYAKE-STONER, S. & MAYFIELD, S. 2010. Micro-algae come of age as a platform for recombinant protein production. *Biotechnology Letters,* 32**,** 1373-1383.
STEVENS, F. J. & ARGON, Y. 1999. Protein folding in the ER. *Seminars in Cell & Developmental Biology,* 10**,** 443-454.
SU, D., ZHAO, H. & XIA, H. 2010. Glycosylation-modified erythropoietin with improved half-life and biological activity. *Int J Hematol,* 91**,** 238-44.
SUZEK, B. E., HUANG, H., MCGARVEY, P., MAZUMDER, R. & WU, C. H. 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics,* 23**,** 1282-1288.
TAN, Z., SHANG, S. & DANISHEFSKY, S. J. 2011. Rational development of a strategy for modifying the aggregatibility of proteins. *Proceedings of the National Academy of Sciences,* 108**,** 4297-4302.
TARTAGLIA, G. G. & VENDRUSCOLO, M. 2008. The Zyggregator method for predicting protein aggregation propensities. *Chemical Society Reviews,* 37**,** 1395-1401.
TETER, S. A., HOURY, W. A., ANG, D., TRADLER, T., ROCKABRAND, D., FISCHER, G., BLUM, P., GEORGOPOULOS, C. & HARTL, F. U. 1999. Polypeptide Flux through Bacterial Hsp70: DnaK Cooperates with Trigger Factor in Chaperoning Nascent Chains. *Cell,* 97**,** 755-765.
THELANDER, L. 1967. Thioredoxin Reductase: CHARACTERIZATION OF A HOMOGENEOUS PREPARATION FROM ESCHERICHIA COLI B. *Journal of Biological Chemistry,* 242**,** 852-859.
THOMAS, J. D., DANIEL, R. A., ERRINGTON, J. & ROBINSON, C. 2001. Export of active green fluorescent protein to the periplasm by the twin-arginine translocase (Tat) pathway in Escherichia coli. *Molecular Microbiology,* 39**,** 47-53.
THOMAS, P. & SMART, T. G. 2005. HEK293 cell line: A vehicle for the expression of recombinant proteins. *Journal of Pharmacological and Toxicological Methods,* 51**,** 187-200.
TIGGES, M. & FUSSENEGGER, M. 2006. Xbp1-based engineering of secretory capacity enhances the productivity of Chinese hamster ovary cells. *Metab Eng*.
TOKURIKI, N., STRICHER, F., SERRANO, L. & TAWFIK, D. S. 2008. How Protein Stability and New Functions Trade Off. *PLoS Comput Biol,* 4**,** e1000002.
TREVINO, S. R., SCHOLTZ, J. M. & PACE, C. N. 2007. Amino acid contribution to protein solubility: Asp, Glu, and Ser contribute more favorably than the other hydrophilic amino acids in RNase Sa. *J Mol Biol,* 366**,** 449-60.
TREVINO, S. R., SCHOLTZ, J. M. & PACE, C. N. 2008. Measuring and increasing protein solubility. *J Pharm Sci,* 97**,** 4155-66.
TROMBETTA, E. S. & PARODI, A. J. 2003. QUALITY CONTROL AND PROTEIN FOLDING IN THE SECRETORY PATHWAY. *Annual Review of Cell and Developmental Biology,* 19**,** 649-676.
TROVATO, A., SENO, F. & TOSATTO, S. C. E. 2007. The PASTA server for protein aggregation prediction. *Protein Engineering Design and Selection,* 20**,** 521-523.

UMA, S., HARTSON, S. D., CHEN, J.-J. & MATTS, R. L. 1997. Hsp90 Is Obligatory for the Heme-regulated eIF-2α Kinase to Acquire and Maintain an Activable Conformation. *Journal of Biological Chemistry,* 272**,** 11648-11656.

VENTURA, S. 2005. Sequence determinants of protein aggregation: tools to increase protein solubility. *Microbial Cell Factories,* 4**,** 11-11.

VENTURA, S. & VILLAVERDE, A. 2006. Protein quality in bacterial inclusion bodies. *Trends in Biotechnology,* 24**,** 179-185.

VOYNOV, V., CHENNAMSETTY, N., KAYSER, V., HELK, B. & TROUT, B. L. 2009. Predictive tools for stabilization of therapeutic proteins. *mAbs,* 1**,** 580-582.

VOYNOV, V., CHENNAMSETTY, N., KAYSER, V., WALLNY, H.-J., HELK, B. & TROUT, B. L. 2010. Design and Application of Antibody Cysteine Variants. *Bioconjugate Chemistry,* 21**,** 385-392.

WAKABAYASHI, H., GRIFFITHS, A. E. & FAY, P. J. 2009. Combining mutations of charged residues at the A2 domain interface enhances factor VIII stability over single point mutations. *Journal of Thrombosis and Haemostasis,* 7**,** 438-444.

WALSH, G. 2014. Biopharmaceutical benchmarks 2014. *Nat Biotech,* 32**,** 992-1000.

WALSH, G. & JEFFERIS, R. 2006. Post-translational modifications in the context of therapeutic proteins. *Nat Biotech,* 24**,** 1241-1252.

WALSH, I., SENO, F., TOSATTO, S. C. E. & TROVATO, A. 2014. PASTA 2.0: an improved server for protein aggregation prediction. *Nucleic Acids Research,* 42**,** W301-W307.

WALTER, S. & BUCHNER, J. 2002. Molecular Chaperones—Cellular Machines for Protein Folding. *Angewandte Chemie International Edition,* 41**,** 1098-1113.

WANG, S., SAKAI, H. & WIEDMANN, M. 1995. NAC covers ribosome-associated nascent chains thereby forming a protective environment for regions of nascent chains just emerging from the peptidyl transferase center. *The Journal of Cell Biology,* 130**,** 519-528.

WANG, W. 2005. Protein aggregation and its inhibition in biopharmaceutics. *Int J Pharm,* 289**,** 1-30.

WANG, W., SINGH, S. K., LI, N., TOLER, M. R., KING, K. R. & NEMA, S. 2012. Immunogenicity of protein aggregates--concerns and realities. *Int J Pharm,* 431**,** 1-11.

WARWICKER, J., CHARONIS, S. & CURTIS, R. A. 2014. Lysine and Arginine Content of Proteins: Computational Analysis Suggests a New Tool for Solubility Design. *Molecular Pharmaceutics,* 11**,** 294-303.

WATERHOUSE, A. M., PROCTER, J. B., MARTIN, D. M., CLAMP, M. & BARTON, G. J. 2009. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics,* 25**,** 1189-91.

WEICKERT, M. J., DOHERTY, D. H., BEST, E. A. & OLINS, P. O. 1996. Optimization of heterologous protein production in Escherichia coli. *Current Opinion in Biotechnology,* 7**,** 494-499.

WEISS, W. F. T., YOUNG, T. M. & ROBERTS, C. J. 2009. Principles, approaches, and challenges for predicting protein aggregation rates and shelf life. *J Pharm Sci,* 98**,** 1246-77.

WEN, D., BOISSEL, J. P., SHOWERS, M., RUCH, B. C. & BUNN, H. F. 1994. Erythropoietin structure-function relationships. Identification of functionally important domains. *J Biol Chem,* 269**,** 22839-46.

WERNER, R. G., KOPP, K. & SCHLUETER, M. 2007. Glycosylation of therapeutic proteins in different production systems. *Acta Pædiatrica,* 96**,** 17-22.

WETZEL, R. 1996. For Protein Misassembly, It's the "I" Decade. *Cell,* 86**,** 699-702.

WILKINSON, B. & GILBERT, H. F. 2004. Protein disulfide isomerase. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics,* 1699**,** 35-44.

WILKINSON, D. L. & HARRISON, R. G. 1991. Predicting the solubility of recombinant proteins in Escherichia coli. *Biotechnology (N Y),* 9**,** 443-8.

WILLEY, K. 1999. An elusive role for glycosylation in the structure and function of reproductive hormones. *Human Reproduction Update,* 5**,** 330-355.

WU, S. J., LUO, J., O'NEIL, K. T., KANG, J., LACY, E. R., CANZIANI, G., BAKER, A., HUANG, M., TANG, Q. M., RAJU, T. S., JACOBS, S. A., TEPLYAKOV, A., GILLILAND, G. L. & FENG, Y. 2010. Structure-based engineering of a monoclonal antibody for improved solubility. *Protein Eng Des Sel,* 23**,** 643-51.

WURM, F. M. 2004. Production of recombinant protein therapeutics in cultivated mammalian cells. *Nat Biotech,* 22**,** 1393-1398.

YAMAMOTO, T., TAKANO, N., ISHIWATA, K., OHMURA, M., NAGAHATA, Y., MATSUURA, T., KAMATA, A., SAKAMOTO, K., NAKANISHI, T., KUBO, A., HISHIKI, T. & SUEMATSU, M. 2014. Reduced methylation of PFKFB3 in cancer cells shunts glucose towards the pentose phosphate pathway. *Nat Commun,* 5**,** 3480.

YASUKAWA, T., KANEI-ISHII, C., MAEKAWA, T., FUJIMOTO, J., YAMAMOTO, T. & ISHII, S. 1995. Increase of solubility of foreign proteins in Escherichia coli by coproduction of the bacterial thioredoxin. *J Biol Chem,* 270**,** 25328-31.

YATES, J. L., WARREN, N. & SUGDEN, B. 1985. Stable replication of plasmids derived from Epstein-Barr virus in various mammalian cells. *Nature,* 313**,** 812-815.

YEWDELL, J. W., LACSINA, J. R., RECHSTEINER, M. C. & NICCHITTA, C. V. 2011. Out with the Old, In with the New? Comparing Methods for Measuring Protein Degradation. *Cell biology international,* 35**,** 457-462.

ZHANG, F., BASINSKI, M. B., BEALS, J. M., BRIGGS, S. L., CHURGAY, L. M., CLAWSON, D. K., DIMARCHI, R. D., FURMAN, T. C., HALE, J. E., HSIUNG, H. M., SCHONER, B. E., SMITH, D. P., ZHANG, X. Y., WERY, J.-P. & SCHEVITZ, R. W. 1997. Crystal structure of the obese protein Ieptin-E100. *Nature,* 387**,** 206-209.

ZHANG, K. & KAUFMAN, R. J. 2006. Protein Folding in the Endoplasmic Reticulum and the Unfolded Protein Response. *In:* STARKE, K. & GAESTEL, M. (eds.) *Molecular Chaperones in Health and Disease.* Springer Berlin Heidelberg.

ZHANG, Y. B., HOWITT, J., MCCORKLE, S., LAWRENCE, P., SPRINGER, K. & FREIMUTH, P. 2004. Protein aggregation during overexpression limited by peptide extensions with large net negative charge. *Protein Expr Purif,* 36**,** 207-16.

ZÖLLS, S., TANTIPOLPHAN, R., WIGGENHORN, M., WINTER, G., JISKOOT, W., FRIESS, W. & HAWE, A. 2012. Particles in therapeutic protein formulations, Part 1: Overview of analytical methods. *Journal of Pharmaceutical Sciences,* 101**,** 914-935.

Chapter 9

Appendices

9.1 Appendix 1

This section details methods for protein mutants develop from papers 1 and 2 (Chapter 3 and 4) that were too detailed to be included due to space limitation.

9.1.1 Site-directed mutagenesis (SDM)

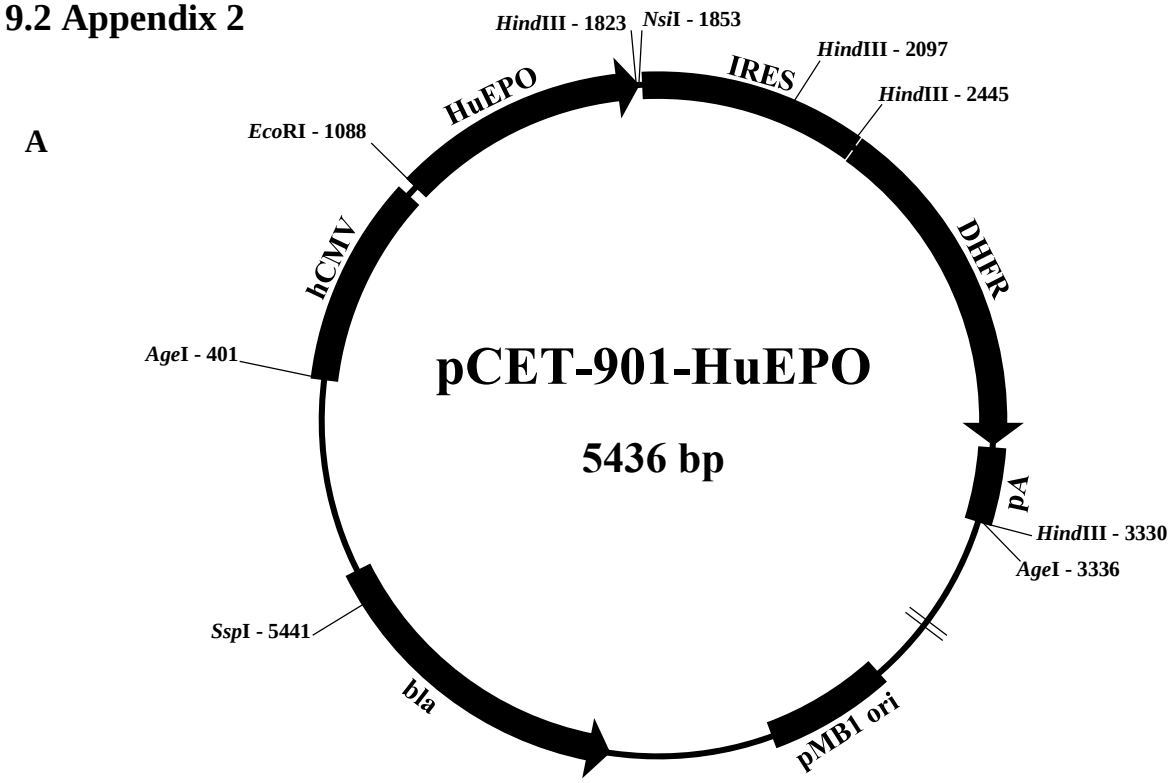
Mutations on rPFKFB3 and rHuEPO protein surfaces were performed by the GENEART Site-Directed Mutagenesis System with the enzyme AccuPrime *Pfx* (Invitrogen). The substitutions were accomplished by enhancing the manufacture's protocol as follows: Mixing 5 µl of 10X AccuPrime *Pfx* Reaction mix, 5 µl of 10X Enhancer, 2.5 µl of each 10 µM primer For and Rev (Appendix Table A.1), 1 µl of 35 ng/µl vector template, 1 µl DNA Methylase (4 U/µl), 2 µl of 25X S-adenosine methionine (SAM), 0.5 µl AccuPrime *Pfx* (2.5 U/µl) and 30.5 µl PCR water. The mutagenesis reaction was carried out with an initial methylation incubation at 37ºC for 20 min followed by a PCR program with an initial denaturation at 95ºC for 2 min, which was followed by 20 cycles consisting of denaturation at 95ºC for 20 s, annealing at 55ºC for 45 s and elongation at 68ºC for 3 or 4 min for rHuEPO or rPFKFB3, respectively. The extra elongation step was at 68ºC for 5 min and the reaction finished at 4ºC. The *in vitro* recombination reaction was done by mixing 4 µl of 5X Reaction Buffer, 10 µl PCR water, 4 µl PCR sample and 2 µl of 10X Enzyme at room temperature for 10 min. The reaction was stopped

by adding 1 µl 0.5 M EDTA and placed it on ice. Then immediately proceeded to *E. coli* cells transformation. 50 µl vial of DH5α^{TM}-T1^{R} *E. coli* cells were thawed on ice for 7 minutes. 2 µl from the recombination reaction were transferred directly into the vial cells and mixed by tapping gently and incubated for 12 min. Then, the reaction tubes were incubated in a water bath for exactly 30 s at 42ºC then put it back on ice for 2 min. 250 µl of pre-warmed SOC medium (Invitrogen) were added to each reaction tube. 100 µL of each samples were transferred to a corresponding LB plate containing ampicillin (100 µg/mL) and spread on the surface. The plates were inverted and incubated at 37ºC for 18-24 h.

Table A9.1. List of oligonucleotides used for SDM

Name	**Oligonucleotide sequence (5' - 3')**	**Description**
rHuEPO		
E13K	For: GACAGCCGAGTCCTG**A**AGAGGTACCTCTTGG Rev: CCAAGAGGTACCTCT**T**CAGGACTCGGCTGTC	Overlapping primers containing the target mutation lysine (AAG)
F48D	For: GACACCAAAGTTAAT**GA**CTATGCCTGGAAGAG Rev: CTCTTCCAGGCATAG**TC**ATTAACTTTGGTGTC	Overlapping primers containing the target mutation aspartic acid (GAC)
R150D	For: GTCTACTCCAATTTCCTC**GAC**GGAAAGCTGAAGCTGTAC Rev: GTACAGCTTCAGCTTTCC**GTC**GAGGAAATTGGAGTAGAC	Overlapping primers containing the target mutation aspartic acid (GAC)
rPFKFB3		
M1	For: CGAGTGCAGAAGATCT**ACAAT**CCC**AA**GGACCACAGGCCCTCG Rev: CGAGGGCCTGTGGTCC**TT**GGG**ATTGT**AGATCTTCTGCACTCG	Overlapping primers containing the target mutations: W13Y, V14N and V16K
M2	For: CATGAAAGTCCGGAAG**G**AATGTGCC**GA**AG**AG**GCCTTGAGAGATGT Rev: ACATCTCTCAAGGC**CT**CT**TC**GGCACATT**C**CTTCCGGACTTTCATG	Overlapping primers containing the target mutations: Q100E, L103E and A104E
M3	For: TTCTTCCGCCCCGAC**GACC**AGGAAGCCATGAAAG Rev: CTTTCATGGCTTCCT**GGTC**GTCGGGGCGAAGAA	Overlapping primers in a two-step PCR containing the target mutations: N91D, E92Q and N178D
	For: GATTACAAAGACTGC**G**ACTCGGCAGAAGCCA Rev: TGGCTTCTGCCGAGT**C**GCAGTCTTTGTAATC	
M4	For: GTCGCTTATGGCTGC**GAC**GTGGAATCCATCTAC Rev: GTAGATGGATTCCAC**GTC**GCAGCCATAAGCGAC	Overlapping primers containing the target mutation aspartic acid (GAC; R427D)

9.2 Appendix 2



B

5' For primer

BamHI

5'atgggggtgcacgaatgtcctgcctggctgtggcttctcctgtccctgctgtcgctccctctgggcctcccagtcctgggcgccccaccacgcctc

Signal peptide sequence — HuEPO

atctgtgacagccgagtcctggagaggtacctcttggaggccaaggaggccgagaatatcacgacgggctgtgctgaacactgcagcttgaatgagaatatc
actgtcccagacaccaaagttaatttctatgcctggaagaggatggaggtcgggcagcaggccgtagaagtctggcagggcctggccctgctgtcggaagct
gtcctgcggggccaggccctgttggtcaactcttcccagccgtgggagcccctgcagctgcatgtggataaagccgtcagtggccttcgcagcctcaccactc
tgcttcgggctctgggagcccagaaggaagccatctcccctccagatgcggcctcagctgctccactccgaacaatcactgctgacactttccgcaaactcttc
cgagtctactccaatttcctccggggaaagctgaagctgtacacaggggaggcctgcaggacaggggacaga**tga**ccaggtg

EcoRI -3'

Stop — Rev primer

C

5' For primer

HindIII

5'atgggggtgcacgaatgtcctgcctggctgtggcttctcctgtccctgctgtcgctccctctgggcctcccagtcctgggcgccccaccacgcctc

Signal peptide sequence — HuEPO

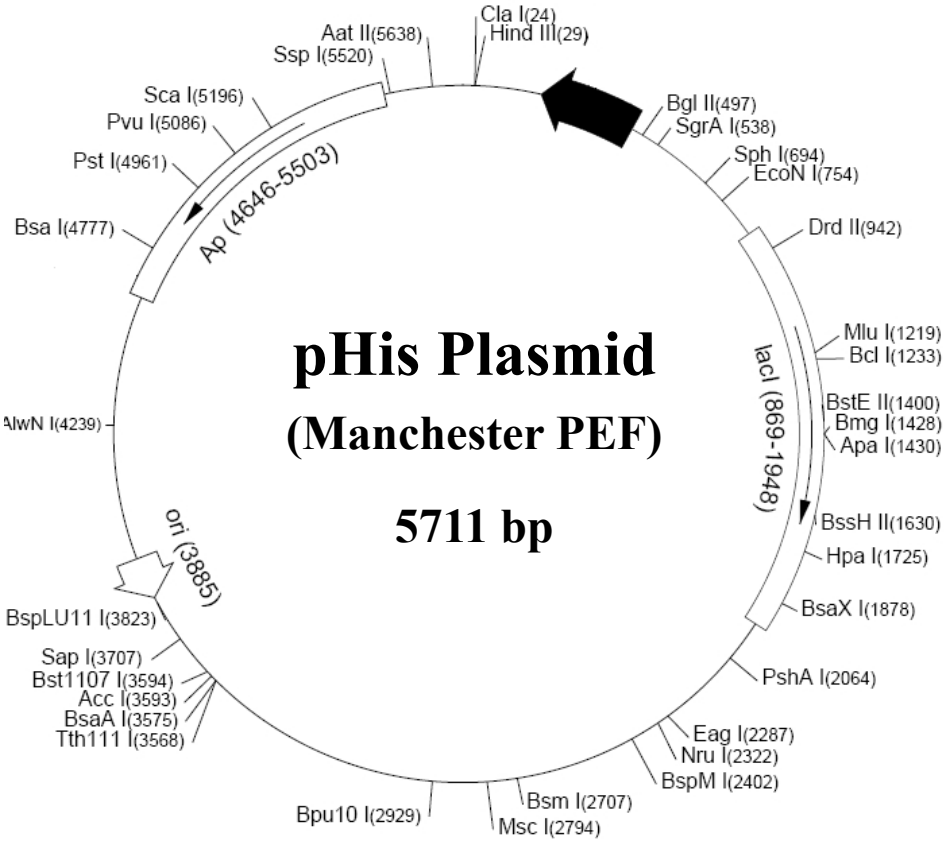
atctgtgacagccgagtcctggagaggtacctcttggaggccaaggaggccgagaatatcacgacgggctgtgctgaacactgcagcttgaatgagaatatc
actgtcccagacaccaaagttaatttctatgcctggaagaggatggaggtcgggcagcaggccgtagaagtctggcagggcctggccctgctgtcggaagct
gtcctgcggggccaggccctgttggtcaactcttcccagccgtgggagcccctgcagctgcatgtggataaagccgtcagtggccttcgcagcctcaccactc
tgcttcgggctctgggagcccagaaggaagccatctcccctccagatgcggcctcagctgctccactccgaacaatcactgctgacactttccgcaaactcttc
cgagtctactccaatttcctccggggaaagctgaagctgtacacaggggaggcctgcaggacaggggacaga

6xHis tag – STOP - 3' Rev 1

NotI -3' Rev 2

Figure A9.1. pCET-901-HuEPO plasmid template vector. (**A**) Full map of pCET-901-HuEPO construct. (**B**) Specific sequence amplification for further insertion into the bacterial pHis expression vector or (**C**) mammalian pCEP-PU expression plasmid.

9.3 Appendix 3



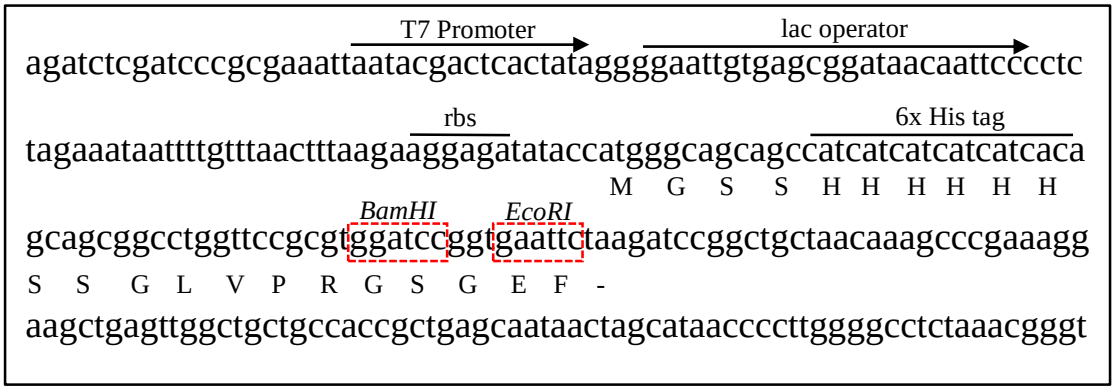


Figure A9.2. Plasmid map and multi cloning site of the pHis vector. This expression plasmid is a modified version of the commercial pET-16b vector (Novagen) encoding a thrombin-cleavable (LVPRGS) amino-terminal 6x His tag. pHis expression vector was developed at the Manchester Protein Expression Facility of the University of Manchester.

9.4 Appendix 4

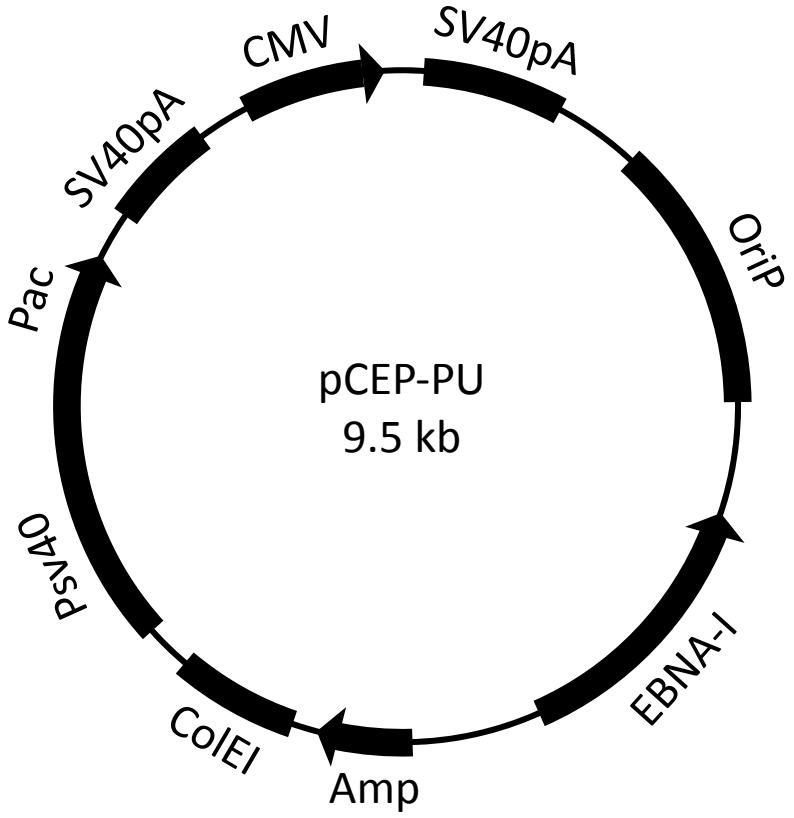


Figure A9.3. pCEP-PU mammalian expression vector. Plasmid map of the transiently transfected recombinant EmGFP and rHuEPO in HEK 293-EBNA cells.