FTB Food Technology & Biotechnology

preliminary communication
ISSN 1330-9862
doi: 10.17113/ftb.56.01.18.5445

# Correlation Between Protein Primary Structure and Soluble Expression Level of HSA dAb in *Escherichia coli*

Yankun Yang[1,2], Guoqiang Liu[1,2], Meng Liu[2], Zhonghu Bai[2,3], Xiuxia Liu[2,3], Xiaofeng Dai[2,3]* and Wenwen Guo[3,4]*

[1] The Key Laboratory of Carbohydrate Chemistry and Biotechnology, School of Biotechnology, Jiangnan University, Ministry of Education, 1800 Lihu Avenue, 214122 Wuxi, PR China
[2] National Engineering Laboratory for Cereal Fermentation Technology, Jiangnan University, 1800 Lihu Avenue, 214122 Wuxi, PR China
[3] Jiangsu Provincial Research Center for Bioactive Product Processing Technology, Jiangnan University, 1800 Lihu Avenue, 214122 Wuxi, PR China
[4] The Key Laboratory of Industrial Biotechnology, Ministry of Education, School of Biotechnology, Jiangnan University, 1800 Lihu Avenue, 214122 Wuxi, PR China

**Received:** July 29, 2017
**Accepted:** January 31, 2018

*Corresponding authors:
Phone/Fax: +8651085329306;
E-mail: xiaofeng.dai@jiangnan.edu.cn (Dai), wendy_gww@foxmail.com (Guo)

ORCID IDs: 0000-0003-4815-6832 (Yang), 0000-0002-9754-587X (Liu, G), 0000-0002-8815-7429 (Liu,M), 0000--0003-3985-7434 (Bai), 0000-0003-1174--2897 (Liu, X), 0000-0002-0006-4042 (Dai), 0000-0002-7145-0303 (Guo)

Paper was presented at the 7th International Forum on Industrial Bioprocessing - IFIBiop 2017, May 21-24, 2017, Wuxi, PR China

## SUMMARY

It is widely accepted that features such as pI, length, molecular mass and amino acid (AA) sequence have a significant influence on protein solubility. Here, we mainly focused on AA composition and explored those that most affected the soluble expression level of human serum albumin (HSA) domain antibody (dAb). The soluble expression and sequence of 65 dAb variants were analysed using clustering and linear modelling. Certain AAs significantly affected the soluble expression level of dAb, with the specific AA combinations being (S, R, N, D, Q), (G, R, C, N, S) and (R, S, G); these combinations respectively affected the dAb expression level in the broth supernatant, the level in the pellet lysate and total soluble dAb. Among the 20 AAs, R displayed a negative influence on the soluble expression level, whereas G and S showed positive effects. A linear model was built to predict the soluble expression level from the sequence; this model had a prediction accuracy of 80 %. In summary, increasing the content of polar AAs, especially G and S, and decreasing the content of R, was helpful to improve the soluble expression level of HSA dAb.

**Key words:** domain antibody (dAb), *Escherichia coli*, heterologous protein soluble expression, linear modelling, primary structure

## INTRODUCTION

Given the outstanding advantages of *Escherichia coli*, including fast growth, inexpensive culturing, high-density cultivation, and simple genetic manipulation, it has been suggested that *E. coli* should be the first host tried for expression of any protein (*1*). However, most proteins from eukaryotes have low solubility when expressed in *E. coli*. For instance, over 80 % of non-membrane proteins were unsuitable for structural studies and over 90 % of potential pharmaceutical proteins were terminated at an early stage of clinical development because of their low solubility when expressed in *E. coli* (*2*). Several strategies have been used to increase protein production and solubility, for example altering expression system elements (*3,4*) and optimizing culture conditions (*5*). These efforts are time-consuming, costly and usually difficult (*6*) because of a lack of understanding of the correlation between the effect of the expression system components and the characteristics of the expressed protein.

Interestingly, it has been found that primary structure features have a great impact on protein overexpression in *E. coli* (*7,8*). Several prediction models have been established (*6,9*), such as the Harrison prediction model (*10*), multiple linear regression (MLR) model (*11*), solubility index-based model (*12*), support vector machine-based model (*13,14*), PROSO model (*15*), SOLpro model (*16*), cc SOL model (*17*) and PROSO II model (*18*). These bioinformatics models can significantly reduce trial and error procedures involved in optimization of expression systems to increase the soluble expression level of heterologous proteins. However, there has been limited application of these prediction models, partly because of the significant differences among the proteins chosen for building them and also because of the adoption of inconsistent culture conditions for expression of proteins (*6,8,9*).

Domain antibodies (dAbs), which consist of only variable regions of heavy ($V_H$) or light ($V_L$) chains (*19*), have simple tertiary structures (Fig. 1; *20,21*), thus it is helpful to focus on

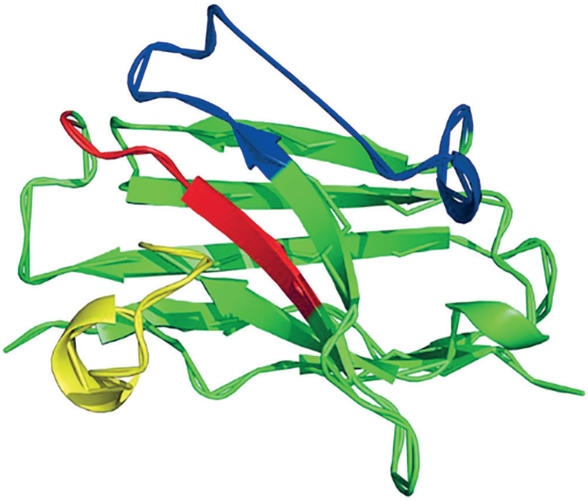

**Fig. 1.** 3D structure of human serum albumin (HSA) domain antibody (dAb) used in this study. 3D structure was obtained from SWISS-MODEL (*20*) and complementarity-determining regions (CDRs) were decorated in three different colours (red, yellow and blue) by PyMOL (*21*)

the features that influence dAb expression level on primary structures. There are three hypervariable regions in dAbs, namely complementarity-determining regions (CDRs) I, II and III, where sequence variability is concentrated to determine the antigen-binding activity of an antibody (*22*). Small variations of amino acids (AAs) within a short region leading to clear variation in soluble expression level, ease of expression in *E. coli* (*23*), and a simple tertiary structure make dAbs an ideal model molecule to investigate the connections between primary structure features and the corresponding soluble protein expression levels.

In this study, a single expression system was used to express multiple human serum albumin (HSA) dAb variants with identical culture and detection conditions, to ensure that no other factors such as culture conditions affect the dAb expression. Clustering and stepwise regression were used to explore the correlation between AA sequences and soluble expression levels of HSA dAbs, aiming at building a linear regression model to predict the soluble expression level of HSA dAb based simply on its AA sequence. Such a model may act as a general guide for site-directed mutagenesis of HSA dAbs or other similar dAbs/Abs to improve the soluble expression levels, which benefits further studies such as interaction mechanism and structure research.

## MATERIALS AND METHODS

### *Random mutation of AAs in the CDRs of the original HSA dAb*

Five amino acids (AAs) were chosen in each complementarity-determining region (CDR) (there are three CDRs, so 15 AAs in total were chosen) to mutate randomly into other AAs, in this way we generated a mutation library consisting of about $10^7$ samples. These samples varied little in pI and molecular mass and had the same length, thus it was helpful to focus on the variables of AA composition. Then, 65 mutated HSA dAbs excluding terminator mutants (AUA, CCU, CCC, AGA and AGG) or sequential repeat mutants were chosen randomly as experimental subjects and 10 were chosen as verification subjects. These mutated sequences are listed in Table 1.

### *Production of recombinant dAb expressing E. coli strains*

The dAb fragments were cloned into vector pBY (an efficient expression vector constructed by a coworker in our lab) and introduced into *E. coli* strain BL21(DE3). The transformed cells were plated onto Luria-Bertani (LB) agar plates (Solarbio® Life Sciences, Beijing, PR China) and incubated at 37 °C overnight. After that, single colonies were selected and inoculated into 25 mL of LB medium (containing 15 μg/mL of tetracycline (Shanghai Shenggong Co. Ltd., Shanghai, PR China) in 250-mL flasks and incubated at 37 °C for 7 h with shaking at 230 rpm. Stock solutions were prepared by mixing 500 μL of culture with 500 μL of 20 % glycerol (Shanghai Hushi Laboratorial Equipment Co. Ltd., Shanghai, PR China) solution in 1.5-mL tubes, and the cells were stored at –80 °C.

### *Cultivation of E. coli strains*

Cultivation can be divided into three phases: seed culture, growth and induction phase. Forty-eight square multititer plates (48-MTP; Thermo Fisher Scientific, Shanghai, PR China) were used to culture the 66 strains (65 mutated strains and a control strain) to achieve parallel fermentation. In the seed culture phase, 2 mL of LB medium containing 15 μg/mL of tetracycline were added into each well of the 48-MTP. After inoculation with 20 μL of stock cell solution, 48-MTPs were incubated in a shaker at 230 rpm and 30 °C for 16 h. In the growth phase, the seed solutions were transferred to fresh 48-MTPs containing 2 mL of Terrific Broth/Super Broth (TB/SB; Solarbio® Life Sciences) medium with 15 μg/mL of tetracycline and cultured under the same conditions as described above. The inoculum volume was calculated by the following equation, thus fixing the initial $A_{595\ \mathrm{nm}}$ at 0.05:

$$V(\mathrm{inoculum})=(0.05\cdot V(\mathrm{fermentation})/A(\mathrm{seed\ culture}))/\mathrm{mL} \quad /1/$$

where *V* is the volume, 0.05 is the initial absorbance (*A*) at 595 nm and *A* is the absorbance of seed culture solution.

Seven hours after the second inoculation, isopropyl-β-D-thiogalactoside (IPTG; Solarbio® Life Sciences) was added to each well to a final concentration of 0.1 mM and the culture temperature was lowered to 23 °C simultaneously. The induction phase lasted for 16 h. After centrifugation of the culture broth at 6000×*g* (centrifuge model Sorvall ST 16R; Thermo Fisher Scientific, Shanghai, PR China), the supernatants were collected, the cell pellets were resuspended in phosphate-buffered saline (PBS; Shanghai Hushi

**Table 1.** Mutation results of 15 animo acids in complementarity-determining regions (CDRs) of human serum albumin (HSA) domain antibodies (dAb)

| No. | CDR1 | CDR2 | CDR3 | No. | CDR1 | CDR2 | CDR3 |
|---|---|---|---|---|---|---|---|
| 1 | SQPHA | WLE-K | QFKHS | 34 | QYKDG | LLTHN | RNGAN |
| 2 | KGNLR | CCSLR | PASTS | 35 | RTPQM | WNVNV | KGGVL |
| 3 | RQYCP | AGVST | T-FMG | 36 | QQLTL | QSWTL | FCALL |
| 4 | ----- | IAYSA | QFYWE | 37 | -KNKP | RRASI | PVSGN |
| 5 | ISNHW | ERVSN | QKFGV | 38 | LPKRL | GFLWI | NKLWQ |
| 6 | YTPLY | FWR-Y | MHLML | 39 | DPREP | VMVKW | P-YDV |
| 7 | PKFCL | SFEGG | KDNYL | 40 | NNNRR | PRYLF | NLHSA |
| 8 | SRCVH | SPA-G | NNYHK | 41 | LDKNA | VLLIC | FGWPV |
| 9 | RGPLS | WTTVL | DK-FT | 42 | DVCFK | LT-AS | WATSN |
| 10 | SYIVP | RAVL- | NLGYL | 43 | WTLCS | VDTAR | FL-RS |
| 11 | -PRHL | CGMTS | WGISP | 44 | ISKST | IPYCQ | NILQL |
| 12 | TVPYR | ALTIG | -KSMS | 45 | KYHQS | RLLLE | KLTLL |
| 13 | PSSIY | CCVDV | WRYEA | 46 | TIWKY | GFVLC | QINEK |
| 14 | RLCPY | NSLGL | SRCHY | 47 | SVGAD | VSVAP | STR-N |
| 15 | TP-VT | VSQ-Q | KTGPL | 48 | YDIGH | QRSRR | AADSD |
| 16 | RWSFR | RTTQN | VNPMR | 49 | ----- | ----- | KLQCT |
| 17 | SGLPT | FTWLI | ETPAL | 50 | GGLSL | GWLTT | IMT-K |
| 18 | ----- | VNG-T | QFTGS | 51 | RANYN | RLGAA | HNMLQ |
| 19 | YYLFS | EFIR- | SCALA | 52 | TAVT- | TA-LP | DEPMR |
| 20 | -RPGL | ASALA | SAVRA | 53 | QL-F- | SWLAS | VDRAA |
| 21 | QNRWL | -GLSS | -K-CP | 54 | EASPR | VNVVP | GLNMR |
| 22 | NTPFL | GNGLV | VNNNN | 55 | GA-VG | ----- | GSVCN |
| 23 | FVITQ | MLRQT | -AYVA | 56 | SQSSQ | PFLFF | CYLPL |
| 24 | AVGTW | DDARS | MAQLA | 57 | CRLTC | LRLQH | VNLQE |
| 25 | AHNAE | PLSLP | SMSCF | 58 | NRNTG | GFLWI | NKLWQ |
| 26 | SILTG | QNCWC | -RNHA | 59 | WCEPS | SAAQS | NSFFE |
| 27 | VPHGG | FRRVN | RVSSK | 60 | ALGCC | FHDSR | SQNTV |
| 28 | TIQQA | CDL-T | VCTGW | 61 | -YRHQ | YTFWT | YGCSK |
| 29 | YTPPR | TG--N | SFWNP | 62 | CTKTL | ----- | VLAVM |
| 30 | DIAGN | RV-HL | QRMKK | 63 | GTITQ | GTSTT | -TYLT |
| 31 | TPESR | C-SES | DGQSD | 64 | SHYNQ | APVES | -VNGL |
| 32 | ILFNL | ----- | SCMAS | 65 | NHAVK | -PIYL | KINTP |
| 33 | LRSLE | D-TSV | MMDLW | Ori | HETMV | HIPPD | LPKRG |
| V1 | SRKWC | DF-FT | RVLGW | V6 | PA-YP | AYVES | AAEKH |
| V2 | SLRAD | QCKFL | RWHTA | V7 | SPHEE | CLT-Y | NNRPW |
| V3 | SVEPS | LKMLG | IYQAT | V8 | KVDTR | RHGQL | CLHPT |
| V4 | VTRSG | SGSDS | NIIST | V9 | ------ | ----- | AI-DN |
| V5 | SHN-L | SRQWQ | VDATQ | V10 | YIPLF | GTIRA | TCWLH |

- no alteration of amino acid at that position

Laboratorial Equipment Co. Ltd) and lysed using Precellys 24 (Bertin Technologies, Paris, France), and then supernatants were collected.

The whole process of cultivation was repeated six times; batches with small deviation of dAb production by control strain were chosen for further analysis, and in this way, parallel operations were guaranteed.

### *Detection and quantification of soluble dAb protein and total protein*

Two amounts of soluble expression of dAbs were measured by direct ELISA, *i.e.* soluble dAbs in broth supernatant and in pellet lysate supernatant. Flat-bottomed 96-well plates (Thermo Fisher Scientific) were first coated with 50 µL of supernatant. After blocking with 5 % non-fat milk in PBS with

Tween 20 (PBST; Shanghai Hushi Laboratorial Equipment Co. Ltd), the dAbs were detected using HRP-labelled protein A (Boster Biological Technology Co. Ltd., Beijing, PR China) with the substrate tetramethylbenzidine (Zhengzhou Biocell Biotechnology Co. Ltd., Zhengzhou, PR China). The reactions were stopped by the addition of 100 µL of 2 M sulfuric acid, and the absorbance was measured at 450 nm/620 nm using an EZ Read 800 (Biochrom, Cambridge, UK). The amount of dAb was calculated from a standard curve made using reference sample. Total protein mass fraction was detected using a modified Bradford protein assay kit (Sangon Biotech Co. Ltd., Shanghai, PR China). To avoid the difference caused by different degrees of cell lysis, standardized amounts of dAbs in µg per g of total protein were calculated as follows and used in the data analysis (Table 2):

$$w(\text{total protein})=m(\text{dAb})/m(\text{total protein}) \qquad /2/$$

*Data analysis*

The software package R (*24*) was used to analyze the contributions of factors such as AA composition, dAb charge and polarity on dAb soluble expression level. Factors with $p<0.05$ were considered significant. Categories of AAs based on Vector NTI® (*25*) are listed in Table 3. Two levels of analysis were run, including dividing expression levels into high and low by Clustal Omega (*26*), and identifying the factors that had an effect on the expression level by *t*-test. A linear regression model was constructed, then factors that had a significant influence were removed in turn to identify the most significant ones based on Akaike information criterion (AIC) values (*27*). We used SWISS-MODEL (*20*) to get 3D structure, and PyMOL (*21*) to decorate CDRs in three different colours.

**Table 2.** Soluble expression data of 65 domain antibody (dAbs) variants using clustering and linear modelling

| No. | γ(soluble dAb in supernatant)/(ng/µL) | γ(soluble dAb in pellet)/(ng/µL) | γ(total soluble dAb)/(ng/µL) |
|---|---|---|---|
| 1 | 3.3±0.4 | 4.3±0.7 | 3.5±0.2 |
| 2 | 3.6±0.6 | 5.0±0.3 | 3.9±0.5 |
| 3 | 4.0±0.8 | 4.2±0.2 | 4.0±0.6 |
| 4 | 4.0±0.8 | 4.3±0.4 | 4.1±0.6 |
| 5 | 5.0±1.1 | 3.8±0.2 | 4.6±0.7 |
| 8 | 3.6±0.3 | 2.3±0.3 | 3.1±0.3 |
| 9 | 3.4±0.5 | 4.3±1.0 | 3.6±0.6 |
| 10 | 3.2±0.4 | 4.5±0.8 | 3.5±0.4 |
| 11 | 3.2±0.3 | 3.1±0.9 | 3.1±0.5 |
| 13 | 3.5±0.7 | 4.5±1.1 | 3.7±0.8 |
| 15 | 3.4±0.6 | 2.5±0.5 | 3.0±0.5 |
| 17 | 5.6±0.5 | 4.2±1.1 | 5.0±0.4 |
| 19 | 3.7±0.2 | 4.0±1.0 | 3.8±0.4 |
| 20 | 3.7±0.6 | 3.1±0.6 | 3.5±0.5 |
| 21 | 4.7±0.8 | 2.9±0.4 | 4.1±0.7 |
| 22 | 1.1±0.3 | 1.2±0.2 | 1.1±0.2 |
| 24 | 2.8±0.6 | 4.9±1.0 | 3.2±0.6 |
| 26 | 4.8±1.2 | 3.0±0.4 | 4.3±0.9 |
| 27 | 3.7±0.4 | 1.9±0.2 | 2.9±0.1 |
| 28 | 3.8±0.8 | 3.3±0.8 | 3.6±0.5 |
| 29 | 4.5±0.1 | 2.4±0.7 | 3.8±0.5 |
| 32 | 3.5±0.7 | 4.3±1.2 | 3.7±0.7 |
| 34 | 0.6±0.1 | 0.8±0.1 | 0.7±0.0 |
| 35 | 3.3±0.4 | 1.3±0.3 | 2.6±0.3 |
| 37 | 3.5±0.6 | 3.8±0.3 | 3.6±0.4 |
| 38 | 4.7±0.5 | 4.2±0.9 | 4.5±0.6 |
| 39 | 2.7±0.7 | 2.3±0.6 | 2.5±0.3 |
| 41 | 4.8±0.3 | 4.6±0.9 | 4.8±0.3 |
| 42 | 0.8±0.1 | 0.5±0.1 | 0.6±0.1 |
| 43 | 3.1±0.6 | 3.2±0.5 | 3.1±0.5 |
| 44 | 5.8±1.0 | 3.4±0.5 | 5.0±0.4 |
| 45 | 4.4±0.3 | 1.2±0.2 | 3.1±0.2 |
| 46 | 3.7±0.6 | 2.4±0.5 | 3.2±0.4 |
| 47 | 4.8±1.3 | 2.5±0.5 | 4.1±1.0 |
| 48 | 3.4±0.5 | 3.3±0.9 | 3.3±0.5 |
| 49 | 3.8±0.5 | 2.3±0.7 | 3.2±0.4 |

Table 2. continued

| No. | γ(soluble dAb in supernatant)/(ng/μL) | γ(soluble dAb in pellet)/(ng/μL) | γ(total soluble dAb)/(ng/μL) |
|---|---|---|---|
| 50 | 4.2±0.7 | 3.8±1.1 | 4.1±0.1 |
| 53 | 4.3±0.9 | 3.1±0.6 | 3.9±0.3 |
| 54 | 4.2±0.9 | 1.8±0.4 | 3.5±0.8 |
| 56 | 3.7±0.8 | 2.0±0.7 | 3.1±0.8 |
| 57 | 3.1±0.3 | 2.4±0.3 | 2.8±0.2 |
| 58 | 4.1±0.8 | 2.5±0.7 | 3.6±0.5 |
| 59 | 3.4±0.6 | 1.8±0.5 | 2.7±0.4 |
| 60 | 4.5±0.5 | 3.9±1.3 | 4.3±0.3 |
| 62 | 2.1±0.3 | 1.4±0.1 | 1.8±0.1 |
| 63 | 4.5±0.7 | 3.2±0.4 | 4.1±0.6 |
| 66 | 2.9±0.2 | 2.7±0.4 | 2.8±0.2 |
| 67 | 4.8±0.9 | 2.8±0.5 | 4.1±0.8 |
| 68 | 5.0±0.7 | 3.1±0.6 | 4.4±0.3 |
| 70 | 0.9±0.1 | 1.3±0.2 | 1.1±0.1 |
| 71 | 4.5±0.1 | 3.8±0.6 | 4.3±0.2 |
| 72 | 2.7±0.4 | 1.5±0.2 | 2.4±0.2 |
| 73 | 1.0±0.2 | 1.2±0.0 | 1.1±0.1 |
| 74 | 3.6±0.5 | 2.0±0.4 | 3.1±0.4 |
| 78 | 5.9±0.9 | 4.0±0.2 | 5.4±0.6 |
| 80 | 5.5±0.9 | 3.2±0.5 | 4.8±0.8 |
| 81 | 1.4±0.3 | 1.3±0.3 | 1.4±0.3 |
| 82 | 3.6±0.6 | 4.0±0.1 | 3.7±0.4 |
| 83 | 4.5±0.9 | 3.3±0.7 | 4.2±0.7 |
| 85 | 3.9±0.9 | 3.6±0.5 | 3.8±0.8 |
| 87 | 2.9±0.4 | 2.6±0.7 | 2.8±0.4 |
| 90 | 3.0±0.3 | 2.9±0.2 | 3.0±0.2 |
| 91 | 4.2±0.9 | 2.2±0.2 | 3.4±0.5 |
| 93 | 5.0±0.6 | 3.1±0.8 | 4.4±0.3 |
| 94 | 3.6±0.7 | 2.4±0.3 | 3.1±0.5 |
| Soluble expression of dAb in validation strains | | | |
| 98 | 1.5±0.2 | 3.3±0.3 | 2.5±0.3 |
| 99 | 3.1±0.2 | 3.5±0.3 | 3.3±0.2 |
| 101 | 3.1±0.2 | 3.3±0.5 | 3.2±0.1 |
| 102 | 3.2±0.4 | 3.5±0.3 | 3.4±0.3 |
| 104 | 3.8±0.5 | 2.9±0.5 | 3.4±0.0 |
| 105 | 3.1±0.2 | 4.0±0.8 | 3.5±0.4 |
| 107 | 2.7±0.3 | 2.9±0.7 | 2.8±0.1 |
| 108 | 3.0±0.2 | 3.0±0.3 | 2.8±0.1 |
| 109 | 1.1±0.3 | 3.2±0.3 | 2.2±0.1 |
| 110 | 3.7±0.5 | 2.7±0.5 | 3.2±0.3 |

Results are expressed as mean value±standard deviation

Table 3. Category of amino acids based on Vector NTI® (*25*)

| Category | Amino acid |
|---|---|
| Charged | R, K, D, C, H, Y, E |
| Polar | N, T, C, G, Q, S, Y |
| Hydrophobic | A, V, L, I, F, W |
| Acidic | D, E |
| Basic | K, R |

## RESULTS

### *AA composition significantly affects the soluble expression of dAbs*

It is widely accepted that AA sequence is significantly correlated with protein production, which was also shown in this study through analysis of the consistency of cluster results based on AA sequences and the corresponding soluble expression levels of dAbs (Table 4). AA compositions of the whole dAb (in percentage) were set as variables to explore

Table 4. Clustering results based on amino acid sequences and soluble expression amounts

| No. | Expresion level | Seq. | Consist. | No. | Expresion level | Seq. | Consist. | No. | Expresion level | Seq. | Consist. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | + | 34 | 2 | 2 | + | 62 | 2 | 1 | – |
| 2 | 1 | 1 | + | 35 | 2 | 2 | + | 63 | 1 | 1 | + |
| 3 | 1 | 1 | + | 37 | 1 | 1 | + | 66 | 2 | 2 | + |
| 4 | 1 | 1 | + | 38 | 1 | 1 | + | 67 | 1 | 1 | + |
| 5 | 1 | 1 | + | 39 | 2 | 1 | – | 68 | 1 | 1 | + |
| 8 | 2 | 2 | + | 41 | 1 | 2 | – | 70 | 2 | 2 | + |
| 9 | 1 | 1 | + | 42 | 2 | 2 | + | 71 | 1 | 1 | + |
| 10 | 1 | 1 | + | 43 | 1 | 1 | + | 72 | 2 | 2 | + |
| 11 | 1 | 2 | – | 44 | 1 | 1 | + | 73 | 2 | 2 | + |
| 13 | 1 | 1 | + | 45 | 2 | 1 | – | 74 | 2 | 2 | + |
| 15 | 2 | 1 | – | 46 | 2 | 2 | + | 78 | 1 | 1 | + |
| 17 | 1 | 2 | – | 47 | 1 | 2 | – | 80 | 1 | 1 | + |
| 19 | 1 | 1 | + | 48 | 1 | 2 | – | 81 | 2 | 1 | – |
| 20 | 1 | 1 | + | 49 | 2 | 2 | + | 82 | 1 | 1 | + |
| 21 | 1 | 1 | + | 50 | 1 | 1 | + | 83 | 1 | 1 | + |
| 22 | 2 | 2 | + | 53 | 1 | 1 | + | 85 | 1 | 1 | + |
| 24 | 1 | 1 | + | 54 | 2 | 1 | – | 87 | 2 | 1 | – |
| 26 | 1 | 1 | + | 56 | 2 | 1 | – | 90 | 1 | 1 | + |
| 27 | 2 | 2 | + | 57 | 2 | 2 | + | 91 | 1 | 2 | – |
| 28 | 1 | 2 | – | 58 | 1 | 2 | – | 93 | 1 | 1 | + |
| 29 | 1 | 2 | – | 59 | 2 | 2 | + | 94 | 2 | 1 | – |
| 32 | 1 | 2 | – | 60 | 1 | 1 | + | Ori | 2 | 1 | – |

1 and 2=cluster result of groups 1 and 2 respectively, based on expression levels or sequences of domain antibodies, + and –=consistency and inconsistency of these two cluster results respectively

their effect on the dAb soluble expression level by a stepwise regression analysis, and the results are summarized in Table 5.

Stepwise regression was taken to analyse AA effect on dAb soluble expression level in broth supernatant, in pellet lysate supernatant and total soluble dAb. Results showed that the combination of AAs S, R, N, D, Q, Y, F and G had a significant influence on dAb soluble yield in broth supernatant, with the p-value of 0.002. Specifically, S, N, D and Q had positive effects, with p-values of 0.0006, 0.02, 0.03 and 0.05, respectively, which means that the soluble yield of dAb in broth supernatant increased with increasing content of these AAs. However, R had a negative effect (p=0.001), thus dAb would be more difficult to express in soluble form in broth supernatant with a higher content of R. Moreover, the combined composition of G, R, C, N, S, Y, K and A had a significant effect on dAb soluble yield in the pellet lysate (p=0.002). Again, R showed a significantly negative effect on the soluble expression (p=0.02), while G, C, N and S showed significantly positive effects, with p-values of 0.01, 0.02, 0.03 and 0.03, respectively. When analyzing AA effect on total amount of soluble dAb, the combined composition of R, S, G, N, Y, C, Q and F showed a significant influence (p=0.0007). The most significant AAs were R (negative), S (positive) and G (positive), for which the p-values were 0.0008, 0.006 and 0.03, respectively (Table 2). Additionally, stepwise regression analysis of the features of the dAbs, including charge, polarity, hydrophobicity, acidity and alkalinity,

Table 5. Amino acids (AAs) that have significantly effect on soluble expression levels of human serum albumin (HSA) domain antibody (dAb)

| Location | AA | Effect | Location | AA | Effect | Location | AA | Effect |
|---|---|---|---|---|---|---|---|---|
| Supernatant | G | + | Pellet | S | + | Total | R | – |
|  | R | – |  | R | – |  | S | + |
|  | C | + |  | N | + |  | G | + |
|  | N | + |  | D | + |  | N | / |
|  | S | + |  | Q | + |  | Y | / |
|  | Y | / |  | Y | / |  | C | / |
|  | K | / |  | F | / |  | Q | / |
|  | A | / |  | G | / |  | F | / |

+=positively/negatively correlated, –=negatively correlated, /=no correlation showed that polarity was the most important feature that had a positive influence on dAb soluble yield (p=0.02).

A linear model was built using total soluble dAb yield data:

$$y=-0.5810\cdot R-0.3711\cdot F+0.4278\cdot G+0.4233\cdot S+ +0.2783\cdot C+0.2737\cdot Q+0.2519\cdot Y+0.2267\cdot N \qquad /3/$$

where y indicates the soluble expression score in %, R is arginine, F is phenylalanine, G is glicine, S is serine, C is cysteine, Q is glutamine, Y is tyrosine and N is asparagine.

The higher the score, the higher the soluble expression level of dAb. Clustering results divided the sequences of the 65 experimental subjects and the control dAb into high- and low-expression groups; the score distribution is shown in Fig. 2. Twenty out of 25 dAbs belonging to the low-expression group had a score <2.5, while 31 out of 41 high-expression dAbs had a score >2.5. We conclude that dAbs with a score <2.5 are likely to be expressed at a low level in soluble form and the soluble yield would possibly be <(2.4±0.9) µg/g. On the other hand, dAbs with a score >2.5 are likely to be expressed at a high level in soluble form, with the potential soluble yield higher than (4.0±0.5) µg/g.

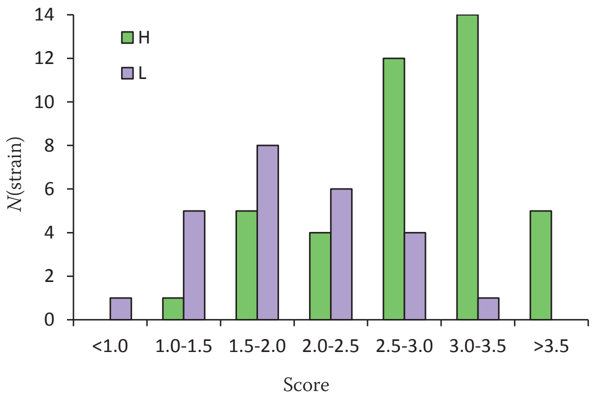

Fig. 2. Score statistics of 66 domain antibodies (dAbs). Twenty out of 25 dAbs belonging to the low-expression group had a score <2.5, while 31 out of 41 high-expression dAbs had a score >2.5

*Verification*

Using the same cultivation and detection methods as in the experiments above, expression data were obtained for 10 verification subjects and a control. Comparing the predicted expression levels from the model with the actual soluble yield of these dAbs, the accuracy of the prediction model was 80 % (Table 6).

## DISCUSSION

Since 1990 there have been many researches exploring the correlation between protein sequence and expression level; however, no consensus has been reached. For example, one project studied 81 different human proteins and came to the conclusion that increasing the average charge, decreasing the number of turn-forming AAs, or decreasing the content of cysteine could reduce the amount of inclusion bodies (*10*), while another studied G-protein-coupled receptors and found that increasing the positive charge encouraged the formation of inclusion bodies (*11*). Goh *et al*. (*28*) discovered that high hydrophobicity was a disadvantage for expressing proteins in soluble form by analyzing 27 267 proteins selected from TargetDB, whereas Luan *et al*. (*29*) expressed 10 167 ORFs of *Caenorhabditis elegans* using a robotic pipeline and found that hydrophobicity was not linearly correlated with the soluble expression level of protein, but proteins with lower hydrophobicity displayed higher levels of soluble expression. These works proved that studies using different subjects could come to different or even opposite conclusions. Here, to avoid the influence of protein properties including molecular mass, length and complex structures, expression system used, or operation bias, first we used dAb as the experimental subject, because this protein has low molecular mass, concentrated regions of variation, is easy to express in *E. coli* and has a simple tertiary structure. Second, 15 AA mutated in CDRs guaranteed enough variation among dAbs and little variation in pI, molecular mass and length, which helped us to focus on the variable of AA composition. Furthermore, we used consistent cultivation conditions and detection methodology to collect data, and repeated the process three times with constant control strain, which guaranteed the parallelity of operation.

Table 6. Comparison between predicted and factual soluble yield of of human serum albumin (HSA) domain antibody (dAb)

| No. | Score | Prediction level | Yield | | Consistency |
|---|---|---|---|---|---|
| | | | *w*/(µg/g) | Level | |
| V1 | 0.5 | Low | 2.5 | Low | Yes |
| V2 | 0.8 | Low | 3.3 | Low | Yes |
| V3 | 3.1 | High | 3.2 | Low | No |
| V4 | 3.9 | High | 3.4 | High | Yes |
| V5 | 2.6 | High | 3.4 | High | Yes |
| V6 | 2.2 | Low | 3.4 | High | No |
| V7 | 2.1 | Low | 2.8 | Low | Yes |
| V8 | 1.1 | Low | 2.8 | Low | Yes |
| V9 | 1.5 | Low | 2.2 | Low | Yes |
| V10 | 1.3 | Low | 3.2 | Low | Yes |

We found that polarity had a significantly positive influence on dAb soluble yield. In other words, the total content of N, S, C, G, T, Q and Y positively correlated with dAb soluble yield. This may be because in this small protein there is a high likelihood of exposure to solvent of polar AAs after folding, which enhances the solubility of the protein through protein–solvent interaction, thus indirectly increasing the soluble expression level of the protein (*30*).

We discovered that arginine content had a significantly negative correlation with dAb soluble yield, consistent with a report that positively charged AAs could hinder the process of translation, thus bringing down the expression level (*7*). Stepwise regression analysis showed that the glycine content was positively correlated with dAb soluble yield, which may be attributable to the small molecular mass and polarity of G. The significantly positive influence of S supports the conclusion that polar AAs benefit dAb soluble expression. We suggest that increasing the total content of G and S, or decreasing the content of R is helpful to improve the soluble expression level of dAb. Findings from this study may act as a general guide for site-directed mutagenesis of HSA dAbs or other similar dAbs/Abs to improve the soluble expression levels, which benefits further studies such as interaction mechanism and structure research. Furthermore, considering the attractive advantages of *E. coli* as a protein expression host, our preliminary observations pave the way towards establishing more efficient *E. coli* expression strategies for desired proteins.

## CONCLUSION

Certain amino acids (AAs) significantly affected the soluble expression level of domain antibody (dAb) in the broth supernatant and in the pellet lysate, and total soluble dAb, with the specific AA combinations being (S, R, N, D, Q), (G, R, C, N, S) and (R, S, G). R displayed a negative influence, whereas G and S showed positive effects. Increasing the content of polar AAs, especially G and S, and decreasing the content of R was helpful to improve the soluble expression level of human serum albumin (HSA) dAb. This linear model had a prediction accuracy of 80 %.

## ACKNOWLEDGEMENTS

We thank all the co-workers who assisted in the screening of mutated sequences, which is indeed a labour-intensive work.

This study was supported by the National Natural Science Foundation of China (grant nos. 31570034 and 31471251), Natural Science Foundation of Jiangsu Province (grant nos. BK20150148 and BK20161130), the Six Talent Peaks Project in Jiangsu Province (grant no. SWYY-128), National Science and Technology Major project (no. 2018ZX10302205-004-002), the 111 Project (no. 111-2-06), the Collaborative Innovation Center of Jiangsu Modern Industrial Fermentation, and the Priority Academic Program Development of Jiangsu Higher Education Institutions.

## REFERENCES

*1.* Gräslund S, Nordlund P, Weigelt J, Hallberg BM, Bray J, Gileadi O, et al. Protein production and purification. Nat Methods. 2008;5(2):135-46.
https://doi.org/10.1038/nmeth.f.202

*2.* Trevino SR, Scholtz JM, Pace CN. 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. 2007;366(2):449-60.
https://doi.org/10.1016/j.jmb.2006.10.026

*3.* Sonoda H, Kumada Y, Katsuda T, Yamaji H. Effects of cytoplasmic and periplasmic chaperones on secretory production of single-chain Fv antibody in Escherichia coli. J Biosci Bioeng. 2011;111(4):465-70.
https://doi.org/10.1016/j.jbiosc.2010.12.015

*4.* Sun W, Xie J, Lin H, Mi S, Li Z, Hua F, Hu Z. A combined strategy improves the solubility of aggregation-prone single-chain variable fragment antibodies. Protein Expr Purif. 2012;83(1):21-9.
https://doi.org/10.1016/j.pep.2012.02.006

*5.* Studier FW. Protein production by auto-induction in high--density shaking cultures. Protein Expr Purif. 2005;41(1):207-34.
https://doi.org/10.1016/j.pep.2005.01.016

*6.* Chang CCH, Song J, Tey BT, Ramanan RN. Bioinformatics approaches for improved recombinant protein production in Escherichia coli: Protein solubility prediction. Brief Bioinform. 2014;15(6):953-62.
https://doi.org/10.1093/bib/bbt057

*7.* Price WN, Handelman SK, Everett JK, Tong SN, Bracic A, Luff JD, et al. Large-scale experimental studies show unexpected amino acid effects on protein expression and solubility in vivo in E. coli. Microb Inform Exp. 2011;1:6.
https://doi.org/10.1186/2042-5783-1-6

*8.* Tian Y, Deutsch C, Krishnamoorthy B. Scoring function to predict solubility mutagenesis. Algorithms Mol Biol. 2010;5:33.
https://doi.org/10.1186/1748-7188-5-33

*9.* Habibi N, Mohd Hashim SZ, Norouzi A, Samian MR. A review of machine learning methods to predict the solubility of overexpressed recombinant proteins in Escherichia coli. BMC Bioinformatics. 2014;15:134.
https://doi.org/10.1186/1471-2105-15-134

*10.* Harrison RG, Bagajewicz MJ. Predicting the solubility of recombinant proteins in Escherichia coli. In: García-Fruitós E, editor. Insoluble proteins, methods in molecular biology (Methods and protocols), vol. 1258. New York, NY, USA: Humana Press; 2015. pp. 403-8.
https://doi.org/10.1007/978-1-4939-2205-5_23

*11.* Kiefer H, Vogel R, Maier K. Bacterial expression of G-protein-coupled receptors: Prediction of expression levels from sequence. Receptors Channels. 2000;7(2):109-19.

*12.* Idicula-Thomas S, Balaji PV. Understanding the relationship between the primary structure of proteins and its propensi-

ty to be soluble on overexpression in Escherichia coli. Protein Sci. 2005;14(3):582-92.
https://doi.org/10.1110/ps.041009005

13. Chan WC, Liang PH, Shih YP, Yang UC, Lin WC, Hsu CN. Learning to predict expression efficacy of vectors in recombinant protein production. BMC Bioinformatics. 2010;11(Suppl 1):S21.
https://doi.org/10.1186/1471-2105-11-S1-S21

14. Idicula-Thomas S, Kulkarni AJ, Kulkarni BD, Jayaraman VK, Balaji PV. A support vector machine-based method for predicting the propensity of a protein to be soluble or to form inclusion body on overexpression in Escherichia coli. Bioinformatics. 2006;22(3):278-84.
https://doi.org/10.1093/bioinformatics/bti810

15. Smialowski P, Martin-Galiano AJ, Mikolajka A, Girschick T, Holak TA, Frishman D. Protein solubility: Sequence based prediction and experimental verification. Bioinformatics. 2007;23(19):2536-42.
https://doi.org/10.1093/bioinformatics/btl623

16. Magnan CN, Randall A, Baldi P. SOLpro: Accurate sequence-based prediction of protein solubility. Bioinformatics. 2009;25(17):2200-7.
https://doi.org/10.1093/bioinformatics/btp386

17. Agostini F, Cirillo D, Livi CM, Delli Ponti R, Tartaglia GG. cc SOL omics: A webserver for solubility prediction of endogenous and heterologous expression in Escherichia coli. Bioinformatics. 2014;30(20):2975-7.
https://doi.org/10.1093/bioinformatics/btu420

18. Smialowski P, Doose G, Torkler P, Kaufmann S, Frishman D. PROSO II - A new method for protein solubility prediction. FEBS J. 2012;279(12):2192-200.
https://doi.org/10.1111/j.1742-4658.2012.08603.x

19. Holt LJ, Herring C, Jespers LS, Woolven BP, Tomlinson IM. Domain antibodies: Proteins for therapy. Trends Biotechnol. 2003;21(11):484-90.
https://doi.org/10.1016/j.tibtech.2003.08.007

20. SWISS-MODEL, Protein Structure Bioinformatics Group, Swiss Institute of Bioinformatics, Biozentrum, University of Basel, Switzerland; 2017. Available from: https://swissmodel.expasy.org/.

21. PyMOL Molecular Graphics System, v. 0.99rc6, Schrödinger, LLC, Cambridge, MA, USA. Available from: https://pymol.org/.

22. Weisser NE Hall JC. Applications of single-chain variable fragment antibodies in therapeutics and diagnostics. Biotechnol Adv. 2009;27(4):502-20.
https://doi.org/10.1016/j.biotechadv.2009.04.004

23. Holt LJ, Basran A, Jones K, Chorlton J, Jespers LS, Brewis ND, Tomlinson IM. Anti-serum albumin domain antibodies for extending the half-lives of short lived drugs. Protein Eng Des Sel. 2008;21(5):283-8.
https://doi.org/10.1093/protein/gzm067

24. R software, v. 3.3.2, Auckland, New Zealand: Statistics Department of the University of Auckland; 2017. Available from: https://www.r-project.org/.

25. Vector NTI®, v. 11.5, Thermo Fisher Scientific, Waltham, MA, USA; 2011. Available from: https://www.thermofisher.com/cn/zh/home/life-science/cloning/vector-nti-software.html.

26. Clustal Omega, Cambridgeshire, UK: EMBL-EBI, European Molecular Biology Laboratory; 2017. Available from: https://www.ebi.ac.uk/Tools/msa/clustalo/.

27. Aho K, Derryberry D, Peterson T. Model selection for ecologists: The worldviews of AIC and BIC. Ecology. 2014;95(3):631-6.
https://doi.org/10.1890/13-1452.1

28. Goh CS, Lan N, Douglas SM, Wu B, Echols N, Smith A, et al. Mining the structural genomics pipeline: Identification of protein properties that affect high-throughput experimental analysis. J Mol Biol. 2004;336(1):115-30.
https://doi.org/10.1016/j.jmb.2003.11.053

29. Luan CH, Qiu SH, Finley JB, Carson M, Gray RJ, Huang W, et al. High-throughput expression of C. elegans proteins. Genome Res. 2004;14:2102-10.
https://doi.org/10.1101/gr.2520504

30. Lesser GJ, Rose GD. Hydrophobicity of amino acid subgroups in proteins. Proteins. 1990;8(1):6-13.
https://doi.org/10.1002/prot.340080104