# 1.Protein–Sol: a web tool for predicting protein solubility from sequence

* https://doi.org/10.1093/bioinformatics/btx345
* 188, 37.6, M. Hebditch+ 3 authorsJ. Warwicker
* 2017, Bioinform.
* Motivation: Protein solubility is an important property in industrial and therapeutic applications. Prediction is a challenge, despite a growing understanding of the relevant physicochemical properties. Results: Protein‐Sol is a web server for predicting protein solubility. Using available data for Escherichia coli protein solubility in a cell‐free expression system, 35 sequence‐based properties are calculated. Feature weights are determined from separation of low and high solubility subsets. The model returns a predicted solubility and an indication of the features which deviate most from average values. Two other properties are profiled in windowed calculation along the sequence: fold propensity, and net segment charge. The utility of these additional features is demonstrated with the example of thioredoxin. Availability and implementation: The Protein‐Sol webserver is available at http://protein‐sol.manchester.ac.uk. Contact: jim.warwicker@manchester.ac.uk

# 2.The CamSol method of rational design of protein mutants with enhanced solubility.

* https://doi.org/10.1016/j.jmb.2014.09.026
* 238, 34.0, P. Sormanni, F. A. Aprile, M. Vendruscolo
* 2015, Journal of molecular biology
* Protein solubility is often an essential requirement in biotechnological and biomedical applications. Great advances in understanding the principles that determine this specific property of proteins have been made during the past decade, in particular concerning the physicochemical characteristics of their constituent amino acids. By exploiting these advances, we present the CamSol method for the rational design of protein variants with enhanced solubility. The method works by performing a rapid computational screening of tens of thousand of mutations to identify those with the greatest impact on the solubility of the target protein while maintaining its native state and biological activity. The application to a single-domain antibody that targets the Alzheimer's Aβ peptide demonstrates that the method predicts with great accuracy solubility changes upon mutation, thus offering a cost-effective strategy to help the production of soluble proteins for academic and industrial purposes.

# 3.DeepSol: a deep learning framework for sequence‐based protein solubility prediction

* https://doi.org/10.1093/bioinformatics/bty166
* 81, 20.25, Sameer Khurana+ 4 authorsRaghvendra Mall
* 2018, Bioinform.
* Motivation: Protein solubility plays a vital role in pharmaceutical research and production yield. For a given protein, the extent of its solubility can represent the quality of its function, and is ultimately defined by its sequence. Thus, it is imperative to develop novel, highly accurate in silico sequence‐based protein solubility predictors. In this work we propose, DeepSol, a novel Deep Learning‐based protein solubility predictor. The backbone of our framework is a convolutional neural network that exploits k‐mer structure and additional sequence and structural features extracted from the protein sequence. Results: DeepSol outperformed all known sequence‐based state‐of‐the‐art solubility prediction methods and attained an accuracy of 0.77 and Matthew's correlation coefficient of 0.55. The superior prediction accuracy of DeepSol allows to screen for sequences with enhanced production capacity and can more reliably predict solubility of novel proteins. Availability and implementation: DeepSol's best performing models and results are publicly deposited at https://doi.org/10.5281/zenodo.1162886 (Khurana and Mall, 2018). Supplementary information: Supplementary data are available at Bioinformatics online.

# 4.SOLpro: accurate sequence-based prediction of protein solubility

* https://doi.org/10.1093/bioinformatics/btp386
* 259, 19.92, C. Magnan, A. Randall, P. Baldi
* 2009, Bioinform.
* MOTIVATION Protein insolubility is a major obstacle for many experimental studies. A sequence-based prediction method able to accurately predict the propensity of a protein to be soluble on overexpression could be used, for instance, to prioritize targets in large-scale proteomics projects and to identify mutations likely to increase the solubility of insoluble proteins. RESULTS Here, we first curate a large, non-redundant and balanced training set of more than 17 000 proteins. Next, we extract and study 23 groups of features computed directly or predicted (e.g. secondary structure) from the primary sequence. The data and the features are used to train a two-stage support vector machine (SVM) architecture. The resulting predictor, SOLpro, is compared directly with existing methods and shows significant improvement according to standard evaluation metrics, with an overall accuracy of over 74% estimated using multiple runs of 10-fold cross-validation.

# 5.Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of Escherichia coli proteins

* https://doi.org/10.1073/pnas.0811922106
* 209, 16.08, Tatsuya Niwa+ 5 authorsH. Taguchi
* 2009, Proceedings of the National Academy of Sciences
* Protein folding often competes with intermolecular aggregation, which in most cases irreversibly impairs protein function, as exemplified by the formation of inclusion bodies. Although it has been empirically determined that some proteins tend to aggregate, the relationship between the protein aggregation propensities and the primary sequences remains poorly understood. Here, we individually synthesized the entire ensemble of Escherichia coli proteins by using an in vitro reconstituted translation system and analyzed the aggregation propensities. Because the reconstituted translation system is chaperone-free, we could evaluate the inherent aggregation propensities of thousands of proteins in a translation-coupled manner. A histogram of the solubilities, based on data from 3,173 translated proteins, revealed a clear bimodal distribution, indicating that the aggregation propensities are not evenly distributed across a continuum. Instead, the proteins can be categorized into 2 groups, soluble and aggregation-prone proteins. The aggregation propensity is most prominently correlated with the structural classification of proteins, implying that the prediction of aggregation propensity requires structural information about the protein.

# 6.PaRSnIP: sequence-based protein solubility prediction using gradient boosting machine

* https://doi.org/10.1093/bioinformatics/btx662
* 62, 15.5, Reda Rawi+ 4 authorsG. Chuang
* 2018, Bioinform.
* Motivation Protein solubility can be a decisive factor in both research and production efficiency, and in silico sequence-based predictors that can accurately estimate solubility outcomes are highly sought. Results In this study, we present a novel approach termed PRotein SolubIlity Predictor (PaRSnIP), which uses a gradient boosting machine algorithm as well as an approximation of sequence and structural features of the protein of interest. Based on an independent test set, PaRSnIP outperformed other state-of-the-art sequence-based methods by more than 9% in accuracy and 0.17 in Matthew's correlation coefficient, with an overall accuracy of 74% and Matthew's correlation coefficient of 0.48. Additionally, PaRSnIP provides importance scores for all features used in training. We observed higher fractions of exposed residues to associate positively with protein solubility and tripeptide stretches with multiple histidines to associate negatively with solubility. The improved prediction accuracy of PaRSnIP should enable it to predict protein solubility with greater reliability and to screen for sequence variants with enhanced manufacturability. Availability and implementation PaRSnIP software is available for download under GitHub (https://github.com/RedaRawi/PaRSnIP). Contact gwo-yu.chuang@nih.gov. Supplementary information Supplementary data are available at Bioinformatics online.

# 7.EPSOL: sequence-based protein solubility prediction using multidimensional embedding

* https://doi.org/10.1093/bioinformatics/btab463
* 15, 15.0, Xiang Wu, Liang Yu
* 2021, Bioinform.
* MOTIVATION The heterologous expression of recombinant protein requires host cells, such as Escherichia coli, and the solubility of protein greatly affects the protein yield. A novel and highly accurate solubility predictor that concurrently improves the production yield and minimizes production cost, and that forecasts protein solubility in an E. coli expression system before the actual experimental work is highly sought. RESULTS In this paper, EPSOL, a novel deep learning architecture for the prediction of protein solubility in an E. coli expression system, which automatically obtains comprehensive protein feature representations using multidimensional embedding, is presented. EPSOL outperformed all existing sequence-based solubility predictors and achieved 0.79 in accuracy and 0.58 in Matthew's correlation coefficient. The higher performance of EPSOL permits large-scale screening for sequence variants with enhanced manufacturability and predicts the solubility of new recombinant proteins in an E. coli expression system with greater reliability. AVAILABILITY AND IMPLEMENTATION EPSOL's best model and results can be downloaded from GitHub (https://github.com/LiangYu-Xidian/EPSOL). SUPPLEMENTARY INFORMATION Supplementary data are available at Bioinformatics online.

# 8.PROSO II – a new method for protein solubility prediction

* https://doi.org/10.1111/j.1742-4658.2012.08603.x
* 142, 14.2, P. Smialowski+ 3 authorsD. Frishman
* 2012, The FEBS journal
* Many fields of science and industry depend on efficient production of active protein using heterologous expression in Escherichia coli. The solubility of proteins upon expression is dependent on their amino acid sequence. Prediction of solubility from sequence is therefore highly valuable. We present a novel machine‐learning‐based model called PROSO II which makes use of new classification methods and growth in experimental data to improve coverage and accuracy of solubility predictions. The classification algorithm is organized as a two‐layered structure in which the output of a primary Parzen window model for sequence similarity and a logistic regression classifier of amino acid k‐mer composition serve as input for a second‐level logistic regression classifier. Compared with previously published research our model is trained on five times more data than used by any other method before (82 000 proteins). When tested on a separate holdout set not used at any point of method development our server attained the best results in comparison with other currently available methods: accuracy 75.4%, Matthew’s correlation coefficient 0.39, sensitivity 0.731, specificity 0.759, gain (soluble) 2.263. In summary, due to utilization of cutting edge machine learning technologies combined with the largest currently available experimental data set the PROSO II server constitutes a substantial improvement in protein solubility predictions. PROSO II is available at http://mips.helmholtz‐muenchen.de/prosoII.

# 9.SOLart: a structure-based method to predict protein solubility and aggregation

* https://doi.org/10.1093/bioinformatics/btz773
* 21, 10.5, Qingzhen Hou+ 2 authorsFabrizio Pucci
* 2020, Bioinform.
* MOTIVATION The solubility of a protein is often decisive for its proper functioning. Lack of solubility is a major bottleneck in high-throughput structural genomic studies and in high-concentration protein production, and the formation of protein aggregates causes a wide variety of diseases. Since solubility measurements are time-consuming and expensive, there is a strong need for solubility prediction tools. RESULTS We have recently introduced solubility-dependent distance potentials that are able to unravel the role of residue-residue interactions in promoting or decreasing protein solubility. Here, we extended their construction by defining solubility-dependent potentials based on backbone torsion angles and solvent accessibility, and integrated them, together with other structure- and sequence-based features, into a random forest model trained on a set of E. coli proteins with experimental structures and solubility values. We thus obtained the SOLart protein solubility predictor, whose most informative features turned out to be folding free energy differences computed from our solubility-dependent statistical potentials. SOLart performances are very good, with a Pearson correlation coefficient between experimental and predicted solubility values of almost 0.7 both in cross validation on the training dataset and in an independent set of S. Cerevisiae proteins. On test sets of modeled structures, only a limited drop in performance is observed. SOLart can thus be used with both high-resolution and low-resolution structures, and clearly outperforms state-of-art solubility predictors. It is available through a user-friendly webserver, which is easy to use by non-expert scientists. AVAILABILITY The SOLart webserver is freely available at http://babylone.ulb.ac.be/SOLART/.

# 10.Protein solubility: sequence based prediction and experimental verification

* https://doi.org/10.1093/bioinformatics/btl623
* 151, 10.07, P. Smialowski+ 4 authorsD. Frishman
* 2007, Bioinform.
* MOTIVATION Obtaining soluble proteins in sufficient concentrations is a recurring limiting factor in various experimental studies. Solubility is an individual trait of proteins which, under a given set of experimental conditions, is determined by their amino acid sequence. Accurate theoretical prediction of solubility from sequence is instrumental for setting priorities on targets in large-scale proteomics projects. RESULTS We present a machine-learning approach called PROSO to assess the chance of a protein to be soluble upon heterologous expression in Escherichia coli based on its amino acid composition. The classification algorithm is organized as a two-layered structure in which the output of primary support vector machine (SVM) classifiers serves as input for a secondary Naive Bayes classifier. Experimental progress information from the TargetDB database as well as previously published datasets were used as the source of training data. In comparison with previously published methods our classification algorithm possesses improved discriminatory capacity characterized by the Matthews Correlation Coefficient (MCC) of 0.434 between predicted and known solubility states and the overall prediction accuracy of 72% (75 and 68% for positive and negative class, respectively). We also provide experimental verification of our predictions using solubility measurements for 31 mutational variants of two different proteins.

# 11.Understanding the relationship between the primary structure of proteins and its propensity to be soluble on overexpression in Escherichia coli

* https://doi.org/10.1110/ps.041009005
* 154, 9.06, S. Idicula-Thomas, P. V. Balaji
* 2005, Protein science : a publication of the Protein Society
* Solubility of proteins on overexpression in Escherichia coli is a manifestation of the net effect of several sequence‐dependent and sequence‐independent factors. This study aims to delineate the relationship between the primary structure and solubility on overexpression. The amino acid sequences of proteins reported to be soluble or to form inclusion bodies on overexpression in E. coli under normal growth conditions were analyzed. The results show a positive correlation between thermostability and solubility of proteins, and an inverse correlation between the in vivo half‐life of proteins and solubility. The amino acid (Asn, Thr, Tyr) composition and the tripeptide frequency of the protein were also found to influence its solubility on overexpression. The amino acids that were seen to be present in a comparatively higher frequency in inclusion body‐forming proteins have a higher sheet propensity, whereas those that are seen more in soluble proteins have a higher helix propensity; this is indicative of a possible correlation between sheet propensity and inclusion body formation. Thus, the present analysis shows that thermostability, in vivo half‐life, Asn, Thr, and Tyr content, and tripeptide composition of a protein are correlated to the propensity of a protein to be soluble on overexpression in E. coli. The precise mechanism by which these properties affect the solubility status of the overexpressed protein remains to be understood.

# 12.Sequence-based prediction of protein solubility.

* https://doi.org/10.1016/j.jmb.2011.12.005
* 89, 8.9, F. Agostini, M. Vendruscolo, G. Tartaglia
* 2012, Journal of molecular biology
* In order to investigate the relationship between the thermodynamics and kinetics of protein aggregation, we compared the solubility of proteins with their aggregation rates. We found a significant correlation between these two quantities by considering a database of protein solubility values measured using an in vitro reconstituted translation system containing about 70% of Escherichia coli proteins. The existence of such correlation suggests that the thermodynamic stability of the native states of proteins relative to the aggregate states is closely linked with the kinetic barriers that separate them. In order to create the possibility of conducting computational studies at the proteome level to investigate further this concept, we developed a method of predicting the solubility of proteins based on their physicochemical properties.

# 13.Improving protein solubility and activity by introducing small peptide tags designed with machine learning models

* https://doi.org/10.1016/j.mec.2020.e00138
* 15, 7.5, Xi Han+ 3 authorsK. Zhou
* 2020, Metabolic engineering communications
* Improving catalytic ability of enzymes is critical to the success of many metabolic engineering projects, but the search space of possible protein mutants is too large to explore exhaustively through experiments. To some extent, highly soluble enzymes tend to exhibit high activity due to their better folding quality. Here, we demonstrate that an optimization algorithm based on a regression model can effectively design short peptide tags to improve solubility of a few model enzymes. Based on the protein sequence information, a support vector regression model we recently developed was used to evaluate protein solubility after small peptide tags were introduced to a target protein. The optimization algorithm guided the sequences of the tags to evolve towards variants that had higher solubility. The optimization results were validated successfully by measuring solubility and activity of the model enzyme with and without the identified tags. The solubility of one protein (tyrosine ammonia lyase) was more than doubled and its activity was improved by 250%. This strategy successfully increased solubility of another two enzymes (aldehyde dehydrogenase and 1-deoxy-D-xylulose-5-phosphate synthase) we tested. The presented optimization methodology thus provides a valuable tool for improving enzyme performance for metabolic engineering and other biotechnology projects.

# 14.Structural hot spots for the solubility of globular proteins

* https://doi.org/10.1038/ncomms10816
* 45, 7.5, A. Ganesan+ 13 authorsJoost Schymkowitz
* 2016, Nature communications
* Natural selection shapes protein solubility to physiological requirements and recombinant applications that require higher protein concentrations are often problematic. This raises the question whether the solubility of natural protein sequences can be improved. We here show an anti-correlation between the number of aggregation prone regions (APRs) in a protein sequence and its solubility, suggesting that mutational suppression of APRs provides a simple strategy to increase protein solubility. We show that mutations at specific positions within a protein structure can act as APR suppressors without affecting protein stability. These hot spots for protein solubility are both structure and sequence dependent but can be computationally predicted. We demonstrate this by reducing the aggregation of human α-galactosidase and protective antigen of Bacillus anthracis through mutation. Our results indicate that many proteins possess hot spots allowing to adapt protein solubility independently of structure and function.

# 15.ccSOL omics: a webserver for solubility prediction of endogenous and heterologous expression in Escherichia coli

* https://doi.org/10.1093/bioinformatics/btu420
* 59, 7.38, Federico Agostini+ 3 authorsGian Gaetano Tartaglia
* 2014, Bioinform.
* Summary: Here we introduce ccSOL omics, a webserver for large-scale calculations of protein solubility. Our method allows (i) proteome-wide predictions; (ii) identification of soluble fragments within each sequences; (iii) exhaustive single-point mutation analysis. Results: Using coil/disorder, hydrophobicity, hydrophilicity, β-sheet and α-helix propensities, we built a predictor of protein solubility. Our approach shows an accuracy of 79% on the training set (36 990 Target Track entries). Validation on three independent sets indicates that ccSOL omics discriminates soluble and insoluble proteins with an accuracy of 74% on 31 760 proteins sharing <30% sequence similarity. Availability and implementation: ccSOL omics can be freely accessed on the web at http://s.tartaglialab.com/page/ccsol\_group. Documentation and tutorial are available at http://s.tartaglialab.com/static\_files/shared/tutorial\_ccsol\_omics.html. Contact: gian.tartaglia@crg.es Supplementary information: Supplementary data are available at Bioinformatics online.

# 16.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

* https://doi.org/10.1093/bioinformatics/bti810
* 102, 6.38, S. Idicula-Thomas+ 3 authorsP. V. Balaji
* 2006, Bioinform.
* MOTIVATION Inclusion body formation has been a major deterrent for overexpression studies since a large number of proteins form insoluble inclusion bodies when overexpressed in Escherichia coli. The formation of inclusion bodies is known to be an outcome of improper protein folding; thus the composition and arrangement of amino acids in the proteins would be a major influencing factor in deciding its aggregation propensity. There is a significant need for a prediction algorithm that would enable the rational identification of both mutants and also the ideal protein candidates for mutations that would confer higher solubility-on-overexpression instead of the presently used trial-and-error procedures. RESULTS Six physicochemical properties together with residue and dipeptide-compositions have been used to develop a support vector machine-based classifier to predict the overexpression status in E.coli. The prediction accuracy is approximately 72% suggesting that it performs reasonably well in predicting the propensity of a protein to be soluble or to form inclusion bodies. The algorithm could also correctly predict the change in solubility for most of the point mutations reported in literature. This algorithm can be a useful tool in screening protein libraries to identify soluble variants of proteins.

# 17.Develop machine learning-based regression predictive models for engineering protein solubility

* https://doi.org/10.1093/bioinformatics/btz294
* 18, 6.0, Xi Han, Xiaonan Wang, Kang Zhou
* 2019, Bioinform.
* MOTIVATION Protein activity is a significant characteristic for recombinant proteins which can be used as biocatalysts. High activity of proteins reduces the cost of biocatalysts. A model that can predict protein activity from amino acid sequence is highly desired, as it aids experimental improvement of proteins. However, only limited data for protein activity are currently available, which prevents the development of such models. Since protein activity and solubility are correlated for some proteins, the publicly available solubility dataset may be adopted to develop models that can predict protein solubility from sequence. The models could serve as a tool to indirectly predict protein activity from sequence. In literature, predicting protein solubility from sequence has been intensively explored, but the predicted solubility represented in binary values from all the developed models was not suitable for guiding experimental designs to improve protein solubility. Here we propose new machine learning models for improving protein solubility in vivo. RESULTS We first implemented a novel approach that predicted protein solubility in continuous numerical values instead of binary ones. After combining it with various machine learning algorithms, we achieved a R2 of 0.4115 when Support Vector Machine (SVM) algorithm was used. Continuous values of solubility are more meaningful in protein engineering, as they enable researchers to choose proteins with higher predicted solubility for experimental validation, while binary values fail to distinguish proteins with the same value - there are only two possible values so many proteins have the same one. AVAILABILITY We present the machine learning workflow as a series of IPython notebooks hosted on GitHub (https://github.com/xiaomizhou616/protein\_solubility). The workflow can be used as a template for analysis of other expression and solubility datasets. SUPPLEMENTARY INFORMATION Supplementary data are available at Bioinformatics online.

# 18.Prediction of protein solubility in Escherichia coli using logistic regression

* https://doi.org/10.1002/bit.22537
* 72, 6.0, Armando A Diaz+ 4 authorsR. Harrison
* 2010, Biotechnology and bioengineering
* In this article we present a new and more accurate model for the prediction of the solubility of proteins overexpressed in the bacterium Escherichia coli. The model uses the statistical technique of logistic regression. To build this model, 32 parameters that could potentially correlate well with solubility were used. In addition, the protein database was expanded compared to those used previously. We tested several different implementations of logistic regression with varied results. The best implementation, which is the one we report, exhibits excellent overall prediction accuracies: 94% for the model and 87% by cross‐validation. For comparison, we also tested discriminant analysis using the same parameters, and we obtained a less accurate prediction (69% cross‐validation accuracy for the stepwise forward plus interactions model). Biotechnol. Bioeng. 2010; 105: 374–383. © 2009 Wiley Periodicals, Inc.

# 19.A review of machine learning methods to predict the solubility of overexpressed recombinant proteins in Escherichia coli

* https://doi.org/10.1186/1471-2105-15-134
* 47, 5.88, N. Habibi+ 2 authorsM. Samian
* 2014, BMC Bioinformatics
* BackgroundOver the last 20 years in biotechnology, the production of recombinant proteins has been a crucial bioprocess in both biopharmaceutical and research arena in terms of human health, scientific impact and economic volume. Although logical strategies of genetic engineering have been established, protein overexpression is still an art. In particular, heterologous expression is often hindered by low level of production and frequent fail due to opaque reasons. The problem is accentuated because there is no generic solution available to enhance heterologous overexpression. For a given protein, the extent of its solubility can indicate the quality of its function. Over 30% of synthesized proteins are not soluble. In certain experimental circumstances, including temperature, expression host, etc., protein solubility is a feature eventually defined by its sequence. Until now, numerous methods based on machine learning are proposed to predict the solubility of protein merely from its amino acid sequence. In spite of the 20 years of research on the matter, no comprehensive review is available on the published methods.ResultsThis paper presents an extensive review of the existing models to predict protein solubility in Escherichia coli recombinant protein overexpression system. The models are investigated and compared regarding the datasets used, features, feature selection methods, machine learning techniques and accuracy of prediction. A discussion on the models is provided at the end.ConclusionsThis study aims to investigate extensively the machine learning based methods to predict recombinant protein solubility, so as to offer a general as well as a detailed understanding for researches in the field. Some of the models present acceptable prediction performances and convenient user interfaces. These models can be considered as valuable tools to predict recombinant protein overexpression results before performing real laboratory experiments, thus saving labour, time and cost.

# 20.Bioinformatics approaches for improved recombinant protein production in Escherichia coli: protein solubility prediction

* https://doi.org/10.1093/bib/bbt057
* 47, 5.88, C. C. Chang+ 2 authorsR. Ramanan
* 2014, Briefings Bioinform.
* The solubility of recombinant protein expressed in Escherichia coli often represents the production yield. However, up-to-date, instances of successful production of soluble recombinant proteins in E. coli expression system with high yield remain scarce. This is mainly due to the difficulties in improving the overall production capacity, as most of the well-established strategies usually involve a series of trial and error steps with unguaranteed success. One way to concurrently improve the production yield and minimize the production cost would be incorporating the potency of bioinformatics tools to conduct in silico studies, which forecasts the outcome before actual experimental work. In this article, we review and compare seven prediction tools available, which predict the solubility of protein expressed in E. coli, using the following criteria: prediction performance, usability, utility, prediction tool development and validation methodologies. This comprehensive review will be a valuable resource for researchers with limited prior experience in bioinformatics tools. As such, this will facilitate their choice of appropriate tools for studies related to enhancement of intracellular recombinant protein production in E. coli.

# 21.Structure-aware protein solubility prediction from sequence through graph convolutional network and predicted contact map

* https://doi.org/10.1186/s13321-021-00488-1
* 5, 5.0, Jianwen Chen+ 2 authorsYuedong Yang
* 2021, Journal of Cheminformatics
* Protein solubility is significant in producing new soluble proteins that can reduce the cost of biocatalysts or therapeutic agents. Therefore, a computational model is highly desired to accurately predict protein solubility from the amino acid sequence. Many methods have been developed, but they are mostly based on the one-dimensional embedding of amino acids that is limited to catch spatially structural information. In this study, we have developed a new structure-aware method GraphSol to predict protein solubility by attentive graph convolutional network (GCN), where the protein topology attribute graph was constructed through predicted contact maps only from the sequence. GraphSol was shown to substantially outperform other sequence-based methods. The model was proven to be stable by consistent $${\text{R}}^{2}$$ R 2 of 0.48 in both the cross-validation and independent test of the eSOL dataset. To our best knowledge, this is the first study to utilize the GCN for sequence-based protein solubility predictions. More importantly, this architecture could be easily extended to other protein prediction tasks requiring a raw protein sequence.

# 22.A relationship between mRNA expression levels and protein solubility in E. coli.

* https://doi.org/10.1016/j.jmb.2009.03.002
* 64, 4.92, G. Tartaglia+ 2 authorsM. Vendruscolo
* 2009, Journal of molecular biology
* Each step in the process of gene expression, from the transcription of DNA into mRNA to the folding and posttranslational modification of proteins, is regulated by complex cellular mechanisms. At the same time, stringent conditions on the physicochemical properties of proteins, and hence on the nature of their amino acids, are imposed by the need to avoid aggregation at the concentrations required for optimal cellular function. A relationship is therefore expected to exist between mRNA expression levels and protein solubility in the cell. By investigating such a relationship, we formulate a method that enables the prediction of the maximal levels of mRNA expression in Escherichia coli with an accuracy of 83% and of the solubility of recombinant human proteins expressed in E. coli with an accuracy of 86%.

# 23.ESPRESSO: A system for estimating protein expression and solubility in protein expression systems

* https://doi.org/10.1002/pmic.201200175
* 39, 4.33, S. Hirose, T. Noguchi
* 2013, Proteomics
* Recombinant protein technology is essential for conducting protein science and using proteins as materials in pharmaceutical or industrial applications. Although obtaining soluble proteins is still a major experimental obstacle, knowledge about protein expression/solubility under standard conditions may increase the efficiency and reduce the cost of proteomics studies.

# 24.Prediction and analysis of protein solubility using a novel scoring card method with dipeptide composition

* https://doi.org/10.1186/1471-2105-13-S17-S3
* 42, 4.2, Hui-Ling Huang+ 8 authorsShinn-Ying Ho
* 2012, BMC Bioinformatics
* BackgroundExisting methods for predicting protein solubility on overexpression in Escherichia coli advance performance by using ensemble classifiers such as two-stage support vector machine (SVM) based classifiers and a number of feature types such as physicochemical properties, amino acid and dipeptide composition, accompanied with feature selection. It is desirable to develop a simple and easily interpretable method for predicting protein solubility, compared to existing complex SVM-based methods.ResultsThis study proposes a novel scoring card method (SCM) by using dipeptide composition only to estimate solubility scores of sequences for predicting protein solubility. SCM calculates the propensities of 400 individual dipeptides to be soluble using statistic discrimination between soluble and insoluble proteins of a training data set. Consequently, the propensity scores of all dipeptides are further optimized using an intelligent genetic algorithm. The solubility score of a sequence is determined by the weighted sum of all propensity scores and dipeptide composition. To evaluate SCM by performance comparisons, four data sets with different sizes and variation degrees of experimental conditions were used. The results show that the simple method SCM with interpretable propensities of dipeptides has promising performance, compared with existing SVM-based ensemble methods with a number of feature types. Furthermore, the propensities of dipeptides and solubility scores of sequences can provide insights to protein solubility. For example, the analysis of dipeptide scores shows high propensity of α-helix structure and thermophilic proteins to be soluble.ConclusionsThe propensities of individual dipeptides to be soluble are varied for proteins under altered experimental conditions. For accurately predicting protein solubility using SCM, it is better to customize the score card of dipeptide propensities by using a training data set under the same specified experimental conditions. The proposed method SCM with solubility scores and dipeptide propensities can be easily applied to the protein function prediction problems that dipeptide composition features play an important role.AvailabilityThe used datasets, source codes of SCM, and supplementary files are available at http://iclab.life.nctu.edu.tw/SCM/.

# 25.ProGAN: Protein solubility generative adversarial nets for data augmentation in DNN framework

* https://doi.org/10.1016/j.compchemeng.2019.106533
* 12, 4.0, Xi Han+ 2 authorsXiaonan Wang
* 2019, Comput. Chem. Eng.
* Abstract Protein solubility plays a critical role in improving production yield of recombinant proteins in biocatalysis applications. To some extent, protein solubility can represent the function and activity of biocatalysts which are mainly composed of recombinant proteins. In literature, many machine learning models have been investigated to predict protein solubility from protein sequence, whereas parameters of those models were underdetermined with insufficient data of protein solubility. Here we propose a deep neural network (DNN) as a more accurate regression predictive model. Moreover, to tackle the insufficient data problem, a novel data augmentation algorithm, Protein Solubility Generative Adversarial Nets (ProGAN), was proposed for improving the prediction of protein solubility. After adding mimic data produced from ProGAN, the prediction performance measured by R2 was improved compared with that without data augmentation. A R2 value of 0.4504 was achieved, which was enhanced about 10% compared with the previous study using the same dataset.

# 26.Using the concept of Chou's pseudo amino acid composition to predict protein solubility: an approach with entropies in information theory.

* https://doi.org/10.1016/j.jtbi.2013.03.010
* 27, 3.0, Niu Xiaohui+ 7 authorsWang Zengzhen
* 2013, Journal of theoretical biology
* Protein solubility plays a major role and has strong implication in the proteomics. Predicting the propensity of a protein to be soluble or to form inclusion body is a fundamental and not fairly resolved problem. In order to predict the protein solubility, almost 10,000 protein sequences were downloaded from NCBI. Then the sequences were eliminated for the high homologous similarity by CD-HIT. Thus, there were 5692 sequences remained. Based on protein sequences, amino acid and dipeptide compositions were generally extracted to predict protein solubility. In this study, the entropy in information theory was introduced as another predictive factor in the model. Experiments involving nine different feature vector combinations, including the above-mentioned three kinds of factors, were conducted with support vector machines (SVMs) as prediction engine. Each combination was evaluated by re-substitution test and 10-fold cross-validation test. According to the evaluation results, the accuracies and Matthew's Correlation Coefficient (MCC) values were boosted by the introduction of the entropy. The best combination was the one with amino acid, dipeptide compositions and their entropies. Its accuracy reached 90.34% and Matthew's Correlation Coefficient (MCC) value was 0.7494 in re-substitution test, while 88.12% and 0.7945 respectively for 10-fold cross-validation. In conclusion, the introduction of the entropy significantly improved the performance of the predictive method.

# 27.Large-scale experimental studies show unexpected amino acid effects on protein expression and solubility in vivo in E. coli

* https://doi.org/10.1186/2042-5783-1-6
* 28, 2.55, W. N. Price+ 11 authorsJ. Hunt
* 2011, Microbial Informatics and Experimentation
* The biochemical and physical factors controlling protein expression level and solubility in vivo remain incompletely characterized. To gain insight into the primary sequence features influencing these outcomes, we performed statistical analyses of results from the high-throughput protein-production pipeline of the Northeast Structural Genomics Consortium. Proteins expressed in E. coli and consistently purified were scored independently for expression and solubility levels. These parameters nonetheless show a very strong positive correlation. We used logistic regressions to determine whether they are systematically influenced by fractional amino acid composition or several bulk sequence parameters including hydrophobicity, sidechain entropy, electrostatic charge, and predicted backbone disorder. Decreasing hydrophobicity correlates with higher expression and solubility levels, but this correlation apparently derives solely from the beneficial effect of three charged amino acids, at least for bacterial proteins. In fact, the three most hydrophobic residues showed very different correlations with solubility level. Leu showed the strongest negative correlation among amino acids, while Ile showed a slightly positive correlation in most data segments. Several other amino acids also had unexpected effects. Notably, Arg correlated with decreased expression and, most surprisingly, solubility of bacterial proteins, an effect only partially attributable to rare codons. However, rare codons did significantly reduce expression despite use of a codon-enhanced strain. Additional analyses suggest that positively but not negatively charged amino acids may reduce translation efficiency in E. coli irrespective of codon usage. While some observed effects may reflect indirect evolutionary correlations, others may reflect basic physicochemical phenomena. We used these results to construct and validate predictors of expression and solubility levels and overall protein usability, and we propose new strategies to be explored for engineering improved protein expression and solubility.

# 28.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.

* https://doi.org/10.1093/jb/mvr042
* 25, 2.27, S. Hirose+ 9 authorsT. Noguchi
* 2011, Journal of biochemistry
* Recombinant protein technology is an important tool in many industrial and pharmacological applications. Although the success rate of obtaining soluble proteins is relatively low, knowledge of protein expression/solubility under 'standard' conditions may increase the efficiency and reduce the cost of proteomics studies. In this study, we conducted a genome-scale experiment to assess the overexpression and the solubility of human full-length cDNA in an in vivo Escherichia coli expression system and a wheat germ cell-free expression system. We evaluated the influences of sequence and structural features on protein expression/solubility in each system and estimated a minimal set of features associated with them. A comparison of the feature sets related to protein expression/solubility in the in vivo Escherichia coli expression system revealed that the structural information was strongly associated with protein expression, rather than protein solubility. Moreover, a significant difference was found in the number of features associated with protein solubility in the two expression systems.

# 29.Discrimination of soluble and aggregation-prone proteins based on sequence information.

* https://doi.org/10.1039/c3mb70033j
* 18, 2.0, Yaping Fang, Jianwen Fang
* 2013, Molecular bioSystems
* Understanding the factors governing protein solubility is a key to grasp the mechanisms of protein solubility and may provide insight into protein aggregation and misfolding related diseases such as Alzheimer's disease. In this work, we attempt to identify factors important to protein solubility using feature selection. Firstly, we calculate 1438 features including physicochemical properties and statistics for each protein. Random Forest algorithm is used to select the most informative and the minimal subset of features based on their predictive performance. A predictive model is built based on 17 selected features. Compared with previous models, our model achieves better performance with a sensitivity of 0.82, specificity 0.85, ACC 0.84, AUC 0.91 and MCC 0.67. Furthermore, a model using a redundancy-reduced dataset (sequence identity <= 30%) achieves the same performance as the model without redundancy reduction. Our results provide not only a reliable model for predicting protein solubility but also a list of features important to protein solubility. The predictive model is implemented as a freely available web application at .

# 30.Predicting the protein solubility by integrating chaos games representation and entropy in information theory

* https://doi.org/10.1016/j.eswa.2013.08.064
* 12, 1.5, Xiaohui Nui+ 3 authorsNana Li
* 2014, Expert Syst. Appl.

# 31.Prediction of protein solubility in E. coli

* https://doi.org/10.1109/eScience.2012.6404416
* 11, 1.1, T. Samak, D. Gunter, Zhong Wang
* 2012, IEEE 8th International Conference on E-Science
* Gene synthesis is a key step to convert digitally predicted proteins to functional proteins. However, it is a relatively expensive and labor-intensive process. About 30-50% of the synthesized proteins are not soluble, thereby further reduces the efficacy of gene synthesis as a method for protein function characterization. Solubility prediction from primary protein sequences holds the promise to dramatically reduce the cost of gene synthesis. This work presents a framework that creates models of solubility from sequence information. From the primary protein sequences of the genes to be synthesized, sequence features can be used to build computational models for solubility. This way, biologists can focus the effort on synthesizing genes that are highly likely to generate soluble proteins. We have developed a framework that employs several machine learning algorithms to model protein solubility. The framework is used to predict protein solubility in the Escherichia coli expression system. The analysis is performed on over 1,600 quantified proteins. The approach successfully predicted the solubility with more than 80% accuracy, and enabled in depth analysis of the most important features affecting solubility. The analysis pipeline is general and can be applied to any set of sequence features to predict any binary measure. The framework also provides the biologist with a comprehensive comparison between different learning algorithms, and insightful feature analysis.

# 32.Deep learning framework DNN with conditional WGAN for protein solubility prediction

* https://www.semanticscholar.org/paper/1c74f5cf382103924685af683bc2a3b9b01cc433
* 2, 0.5, X. Han+ 2 authorsX. Wang
* 2018, arXiv: Quantitative Methods
* Protein solubility plays a critical role in improving production yield of recombinant proteins in biocatalyst and pharmaceutical field. To some extent, protein solubility can represent the function and activity of biocatalysts which are mainly composed of recombinant proteins. Highly soluble proteins are more effective in biocatalytic processes and can reduce the cost of biocatalysts. Screening proteins by experiments in vivo is time-consuming and expensive. In literature, large amounts of machine learning models have been investigated, whereas parameters of those models are underdetermined with insufficient data of protein solubility. A data augmentation algorithm that can enlarge the dataset of protein solubility and improve the performance of prediction model is highly desired, which can alleviate the common issue of insufficient data in biotechnology applications for developing machine learning models. We first implemented a novel approach that a data augmentation algorithm, conditional WGAN was used to improve prediction performance of DNN for protein solubility from protein sequence by generating artificial data. After adding mimic data produced from conditional WGAN, the prediction performance represented by $R^{2}$ was improved compared with the $R^{2}$ without data augmentation. After tuning the hyperparameters of two algorithms and organizing the dataset, we achieved a $R^{2}$ value of $45.04\%$, which enhanced $R^{2}$ about $10\%$ compared with the previous study using the same dataset. Data augmentation opens the door to applications of machine learning models on biological data, as machine learning models always fail to be well trained by small datasets.

# 33.A benchmark of protein solubility prediction methods on UDP-dependent glycosyltransferases

* https://doi.org/10.1101/2020.02.28.962894
* 1, 0.5, F. Ghomi, Tiia Kittilä, D. Welner
* 2020, bioRxiv
* UDP-dependent glycosyltransferases (UGTs) are enzymes that glycosylate a wide variety of natural products, thereby modifying their physico-chemical properties, i.e. solubility, stability, reactivity, and function. To successfully leverage the UGTs in biocatalytic processes, we need to be able to screen and characterise them in vitro, which requires efficient heterologous expression in amenable hosts, preferably Escherichia coli. However, many UGTs are insoluble when expressed in standard and attempted optimised E. coli conditions, resulting in many unproductive and costly experiments. To overcome this limitation, we have investigated the performance of 11 existing solubility predictors on a dataset of 57 UGTs expressed in E. coli. We show that SoluProt outperforms other methods in terms of both threshold-independent and threshold-dependent measures. Among the benchmarked methods, only SoluProt is significantly better than random predictors using both measures. Moreover, we show that SoluProt uses a threshold for separating soluble and insoluble proteins that is optimal for our dataset. Hence, we conclude that using SoluProt to select UGT sequences for in vitro investigation will significantly increase the success rate of soluble expression, thereby minimising cost and enabling efficient characterisation efforts for biocatalysis research.

# 34.Prediction of soluble heterologous protein expression levels in Escherichia coli from sequence-based features and its potential in biopharmaceutical process development

* https://doi.org/10.4155/PBP.14.23
* 1, 0.12, Xiaofeng Dai+ 5 authorsZhonghu Bai
* 2014, Pharmaceutical bioprocessing
* Prediction of soluble protein expression levels in Escherichia coli based on the nature of protein itself remains a challenge for bioprocess development (BD). This review will critically discuss the current efforts and achievements that employ computational approaches to develop prediction models for soluble protein expression in E. coli. The contrast between the remarkable progresses made on the predictive models achieved by bioinformatics and their relatively infrequent application in BD will be explained. The effects of process-relevant variables at four different levels on the expression of heterologous proteins, for example, gene, vector, host cell and cultivation process, and also a critical comparison of several established bioinformatics tools for predicting expression levels will be presented. The potential utility of this emergent technology to increase the efficiency of BD strategies and thereby to reduce the cost of establishing a process for soluble protein expression are critically examined.

# 35.Sequence analysis Protein solubility : sequence based prediction and experimental verification

* https://scholar.google.com/scholar?q=Sequence%20analysis%20Protein%20solubility%20%3A%20sequence%20based%20prediction%20and%20experimental%20verification
* 0, 0.0, P. Smialowski+ 4 authorsD. Frishman
* 2007,
* Motivation: Obtaining soluble proteins in sufficient concentrations is a recurring limiting factor in various experimental studies. Solubility is an individual trait of proteins which, under a given set of experimental conditions, is determined by their amino acid sequence. Accurate theoretical prediction of solubility from sequence is instrumental for setting priorities on targets in large-scale proteomics projects. Results: We present a machine-learning approach called PROSO to assess the chance of a protein to be soluble upon heterologous expression in Escherichia coli based on its amino acid composition. The classification algorithm is organized as a two-layered structure in which the output of primary support vector machine (SVM) classifiers serves as input for a secondary Naive Bayes classifier. Experimental progress information from the TargetDB database aswell as previously published datasets were used as the source of training data. In comparison with previously published methods our classification algorithmpossesses improved discriminatory capacity characterized by the Matthews Correlation Coefficient (MCC) of 0.434 between predicted and known solubility states and the overall prediction accuracy of 72% (75 and 68% for positive and negative class, respectively). We also provide experimental verification of our predictions using solubility measurements for 31 mutational variants of two different proteins. Availability: A web server for protein solubility prediction is available at http://webclu.bio.wzw.tum.de:8080/proso Contact: d.frishman@wzw.tum.de Supplementary information: Supplementary data are available at Bioinformatics online