University of Lausanne

Department of Computational Biology

First-step project

Data extraction and machine learning in features involved in coevolution

Author

Karim Saied

December 2017

Abstract

Coevolution is a fundamental process observed at different levels in nature. At molecular level, several methods are used to detect coevolving pairs. Once done, their distribution is generally investigated in order to highlight their location within and between proteins, which provide useful structural and functional information. In this study, coevolving positions pairs in several human genes were first used to assess the relationship between coevolution and structural features in the genes products. The impact of gene conservation was also evaluated. The results showed that beta strands tend to contain a higher amount of coevolving positions compared to other secondary structures. Moreover, relatively variable proteins seem to display more related pairs than conserved proteins. As a second step, by using machine learning, another approach to investigate coevolution was introduced. As a matter of fact, predictions about coevolution scores and the distributions of positions pairs among the secondary structures have been attempted. Unfortunately, the dataset used by the machine learning algorithm suffered from a lack of information as only two features were used to predict a third one. As a consequence, none of the predictions displayed a sufficient accuracy.

1 Introduction

Coevolution is a process observed in different biological systems and implies a coordinated modification between two elements. As a matter of fact, the modification of an element triggers the change of another element [8]. At species level, a well-known example is the relationship between the length of spur in some orchids and the length of proboscis in some pollinators such as the sphinx moth *Xanthopan morgani.* Indeed, longer spurs select pollinators with longer proboscis. At molecular level, coordinated changes are observed between nucleotides or amino acids pairs, depending whether the process occurs in DNA/RNA molecules or proteins. Due to selective constraints acting on these biomolecules some disadvantageous mutations at specific sites are sometimes compensated by the modification of another position in the same sequence or on another protein. In doing so, a structural and functional integrity of proteins/proteins complexes or RNA molecules is maintained [7].

To quantify covariation between positions pairs in a given sequence, many computational methods have been developed so far. One can site the mutual information model, maximum likelihood estimation, Bayesian probabilities and phylogenetic approaches [3]. A multiple sequence alignment (MSA) of a protein family is generally used as a starting point for a model [3]. Although the evaluation of coevolution begins with the detection of related positions pairs in sequences, further analysis are then required to assess their distribution along a sequence or between proteins. Such step allows to get information about proteins structures, their functions and their interactions.

As some genes products or proteins complexes tend to be maintained functional through coordinated mutations in their sequence, one might wonder about the relationship between coevolution and conservation. In fact, it has been observed that conserved sites within proteins tend to have a higher amount of coevolving positions whereas conserved protein families possess fewer pairs [6]. In other words, a correlation was found between coevolution and conservation. Indeed, the correlation is positive at the level of the protein site whereas the trend is negative at the level of the protein family. Furthermore, the same study highlighted the fact that coordinated mutations are more likely to be present within alpha helices than beta strands or turns. That observation highlights a relationship between coevolution and structural features of proteins as some secondary structures are likely to contain more related pairs.

The current study aims to determine which protein features could explain the amount of coevolution observed in a gene. In other words, the relationship between protein features and coevolution will be evaluated. For that purpose, the previously mentioned observations made by Mandloi *et al.*, 2017 have been used as a starting point and has allowed to starts with relevant protein features to test, namely the secondary structures and the gene conservation. The latter was calculated using the multiple sequence alignments of specific human genes and the structural features of the corresponding proteins were extracted from UniProt. In addition, coevolution predictions about the genes used were collected from CoevDB, a database storing information about coevolution in vertebrate genomes. As a second step, a type of machine learning algorithm called Decision Tree was used to attempt to predict the coevolution scores and the distribution of positions pairs in secondary structures.

2 Results

2.1 Distribution of coevolving positions in secondary structures

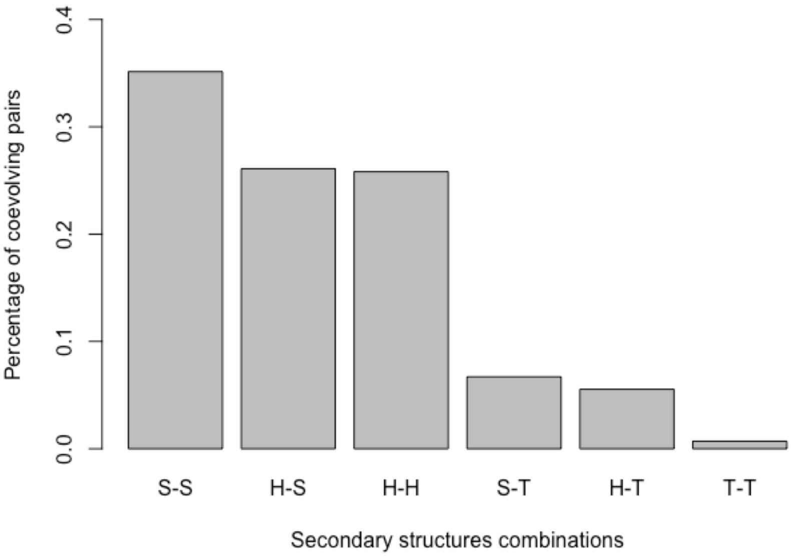

Figure 1: **Distribution of coevolving positions among the different secondary structures.** The letters *H*, *S* and *T* stand for alpha helix, beta strand and turn respectively. The graph displays the amount of coevolving positions that are both located in alpha helices (*H*-*H*), beta strands (*S*-*S*) or turns (*T*-*T*) but also those that are in two different secondary structures (*H*-*S*), (*H*-*T*), (*S*-*T*).

As a first step, the distribution of the coevolving sites in the different secondary structures was assessed. It was found that most coevolving positions were located either in a same beta strand or in two different ones. Such result is not consistent with what has been observed by Mandloi *et al.*, 2017, namely that coevolving sites are more prevalent in alpha helices than in other secondary structures. Indeed, the current results show that 57609 (35.1%) positions pairs displayed the pattern *S-S* compared with 42334 (25.8%) for the pattern *H-H*. The investigation was extended to other secondary structures patterns as well. As a matter of fact, 42776 pairs (26.1%) were associated with the pattern *H-S*, 10986 (6.7%) with *S-T*, 9079 (5.5%) with *H-T* and only 1146 pairs (0.7%) displayed the pattern *T-T*.

2.2 Genes conservation and amount of coevolving pairs

The relationship between the amount of coevolving pairs in a gene and its conservation was assessed. One can note a significant negative correlation (p-value: 2.7e-06), although it is relatively weak. Such result is consistent with what has been observed in the literature [6], namely that relatively conserved protein families tend to contain less coevolving sites than relatively variable families.

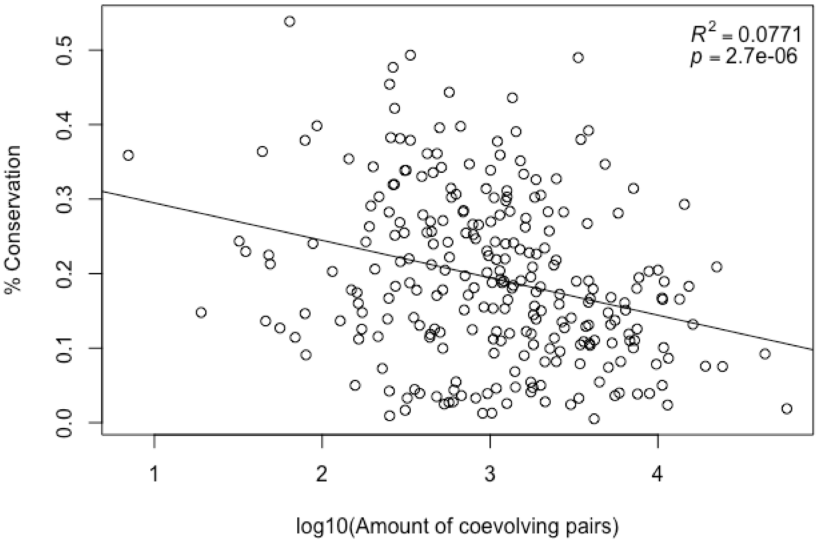

Figure 2: **Conservation of genes according to their amount of coevolving pairs.** Each dot corresponds to a gene with a specific percentage of conservation and containing a certain amount of coevolving positions

2.3 Machine learning to predict coevolution and structural features

By using machine learning, another approach to evaluate coevolution is presented. Two types of Decision Tree were used: a Regression Tree and a Classification Tree. First, the secondary structure patterns and the gene conservation were used to predict the coevolution scores, namely the ΔAIC values. To this end, the Regression Tree was applied. The algorithm predicted the scores with 12.50 % ± 5.96 % accuracy.

Then, the coevolution score and the gene conservation were used to predict the secondary structures patterns. Therefore, the Classification Tree was applied and predictions were made with 45.50 % ± 6.67 % accuracy.

3 Discussion

By using machine learning, the project aimed to use relevant protein features and assess which of them were able to explain the coevolution found in a gene. In that sense, and starting from observations made by Mandloi *et al.*, 2017, the project tried to evaluate whether structural features or conservation were reliable predictors of coevolution. The analysis was therefore structured around structural features and the conservation of human genes.

Regarding the distribution of positions pairs among the different secondary structures (Figure 1), the results were not consistent with what has been found in the literature. One has to note that a sample composed of 448 human genes was used, thus a bigger sample should be used. Regarding the relationship between the gene conservation and the amount of coevolving pairs, the results showed a negative correlation between these elements. As a matter of fact, relatively more conserved protein tend to have less coevolving pairs than relatively variable proteins. That result is quite intuitive because highly conserved proteins are generally related to important/vital mechanisms in the cell. Hence, any mutation of a position in the sequence would probably lead to the death of the organism without allowing another position to compensate the first mutation. Moreover, the probability that two critical sites could undergo two compensatory mutations at the same time is very low. In that sense, one could suggest that comparative analysis of the amount of coevolving pairs in proteins could provide some clue and an overview on proteins conservation.

The Regression Tree was not able to predict ΔAIC values with sufficient accuracy, meaning that either the features used were not reliable predictors of coevolution or the amount of features used was not enough. Knowing that only two features were taken into account as input, namely the secondary structure patterns and the conservation, it is more likely that the dataset suffered from a lack of information. Moreover, the gene conservation value was the same for all the positions pairs in a given gene. As a matter of fact, this element was practically not informative and did not carry that much weight in the prediction made by the model. Instead of using gene conservation, more information would be carried by the site conservation. In that sense, each position pair would have its own conservation value, which would allow them to be independent from each other and, hence, having a specific set of values for each instance. Regarding the Classification Tree, the predictions of secondary structures patterns were made with a relatively higher accuracy but still not sufficient. Here too, the use of the gene conservation provided almost no information for the same reasons given above.

In any case, gene conservation should be kept in the dataset, tough, because it would still provide a certain amount of information to the model, no matter how small. However, it is clear that additional proteins features should be taken into account. As a matter of fact, Mandloi *et al.*, 2017 also investigated the influence of molecular and biological functions as well as the cellular location of proteins. These elements would be relevant and would add a significant added value to the dataset. However, because of the large range of molecular and biological functions carried out by some proteins, it is necessary to find a way to summarize such information.

4 Materials and methods

4.1 Tools

The current project, was performed using Python programming (Python 3.6.2) through the software Spyder (version 3.2.4). The relational database management system MySQL was used to store data about the genes used and data extracted from UniProt.

4.2 Dataset from CoevDB

First, data about 448 human genes were collected from CoevDB, a database containing predictions about coevolution of position pairs in DNA sequences from vertebrates. These predictions were previously made by means of Coev, a probabilistic model developed by Dib *et al.*, 2014.

The dataset extracted contained information about the different human genes used, namely an ID, the coevolving positions pairs and their coevolution score (ΔAIC). The latter measures the strength of coevolution between two positions and was calculated by testing the Coev model against another one that assumes independent changes between sites ($\Delta\text{AIC} = AIC_{independent} - AIC_{Coev}$) [1]. The higher the score, the greater the strength of coevolution. Data have been restricted to coevolving pairs displaying a ΔAIC greater than 20, which was considered, in this study, as a good evidence of coevolution.

4.3 Dataset from UniProt

In addition to data about human genes extracted from CoevDB, information about their products were gathered from UniProt and stored in different MySQL tables. The Universal Protein Resource (UniProt) is a database containing information about protein sequences and related annotations [2]. It is a highly useful resource to access to information about proteins such as their function, structures, amino acids sequences, their length, and so on.

Data retrieval started from a collection of FASTA files, each of them containing the multiple sequence alignment of a gene shared between different vertebrate species. Data extracted were restricted to human genes found in the dataset from CoevDB. The Python script created handled the FASTA files one after the other as described in the following paragraph.

As a first step, for each FASTA file, the header lines (Ensembl ID) of human sequences were collected and used through UniProt to retrieve the corresponding entry as well as the entry name on the website. As an example, the Ensembl ID *ENSP00000376457* is related to the entry *Q9H8V3* and the entry name *ECT2_HUMAN* on UniProt, which corresponds to the epithelial cell-transforming sequence 2 oncogene. Each entry was then used to access the online text file of the corresponding gene product. Each text file was screened in order to determine whether or not the gene product had secondary structures, molecular functions or was involved in some biological processes. If present, the different functions and processes were extracted and stored in two different MySQL tables. The secondary structures along with their range of amino acids were gathered in another table.

4.4 Calculation of gene conservation

Along with data retrieval from CoevDB and UniProt, the conservation of each gene used was calculated using its multiple sequence alignment:

$$conserv = \frac{I}{N} \tag{1}$$

Where I denotes the amount of sites that are fully conserved through a given alignment and N the total number of sites in the sequences.

4.5 Relationship between conservation and the amount of coevolving pairs

To evaluate the relationship between the conservation and the amount of coevolving pairs in the genes used, a linear regression was performed using R (version 3.4.3). It appeared the residuals were normally distributed but heteroscedasticity was observed. To correct that, a log-transformation was performed on the amount of coevolving pairs.

4.6 Decision Tree algorithm and data processing

In order to be used by the machine learning algorithm, the most relevant protein features dispersed across all MySQL tables previously created had to be grouped in a single dataset. The latter was composed of the gene IDs, coevolving positions, their secondary structures pattern, their ΔAIC value and the gene conservation (Table 1). In machine learning, Decision Tree is a type of supervised learning algorithm. It can be used for classification and regression problems [4, 5].

Gene ID	Position 1	Position 2	Secondary structures	Conservation	DAIC
2	371	377	H-T	0.314044	21.2045
2	130	190	H-S	0.314044	28.6688
2	187	366	H-H	0.314044	22.2352
759	177	220	H-S	0.327350	24.5655
759	177	220	H-S	0.327350	35.9264
759	209	252	S-S	0.327350	24.7965
759	209	273	S-S	0.327350	21.2641
759	211	220	H-S	0.327350	21.9099
...	...	...	...	...	...
7325	140	167	H-H	0.100304	32.0210
7325	143	144	H-H	0.100304	20.0833
7325	144	146	H-H	0.100304	24.2946
7325	144	147	H-T	0.100304	24.8935
7325	144	147	H-T	0.100304	22.9442

Table 1: **Part of the Decision Tree dataset.** The gene IDs as well as the positions were not used for the analysis but allowed to keep track of the genes and their respective coevolving positions.

The dataset was splitted into two subsets: 90% of the instances were allocated to the training set and the 10% remaining for the test set. For the Regression Tree, the secondary structures patterns and the conservation were the variables used as input whereas the ΔAIC was the target variable (i.e. the variable to be predicted). Regarding the Classification Tree, the ΔAIC and the gene conservation were used as input variables and the secondary structures patterns was the target variable. To assess the accuracy of the predictions, a 10-fold cross validation was performed.

Both Decision Tree methods used for the analysis came from the Scikit-learn library using the *sklearn* package (version 0.18.1) in Python. The Decision Tree regression and classification were performed using the functions *DecisionTreeRegressor()* and *DecisionTreeClassifier()* respectively. Then, the *fit(x_train, y_train)* method was applied using the training sets as arguments to allow the estimator to learn from the data. Finally, predictions were made by means of the *predict(x_validation)* method, which allowed to return the predicted values Y from the validation set X.

5 Acknowledgements

I would like to thank Rocio Rama Ballesteros and Prof. Nicolas Salamin for their help and advice throughout the project. I also thank Sarah Schmid and Pablo Duchen for their advice concerning my oral defense.

References

[1] Linda Dib, Daniele Silvestro, and Nicolas Salamin. Evolutionary footprint of coevolving positions in genes. *Bioinformatics*, 30(9):1241–1249, 2014.

[2] Bateman et al. UniProt: The universal protein knowledgebase. *Nucleic Acids Research*, 45(D1):D158–D169, 2017.

[3] Hongyun Gao, Yongchao Dou, Jialiang Yang, and Jun Wang. New methods to measure residues coevolution in proteins. *BMC bioinformatics*, 12(1):206, 2011.

[4] Carl Kingsford and Steven L. Salzberg. What are decision trees? *Nature Biotechnology*, 26(9):1011–1012, 2008.

[5] Wei-Yin Loh. Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):14–23, 2011.

[6] Sapan Mandloi and Saikat Chakrabarti. Protein sites with more coevolutionary connections tend to evolve slower, while more variable protein families acquire higher coevolutionary connections. *F1000Research*, 6(0):453, 2017.

[7] Chen Hsiang Yeang and David Haussler. Detecting coevolution in and among protein domains. *PLoS Computational Biology*, 3(11):2122–2134, 2007.

[8] Kevin Y. Yip, Prianka Patel, Philip M. Kim, Donald M. Engelman, Drew Mcdermott, and Mark Gerstein. An integrated system for studying residue coevolution in proteins. *Bioinformatics*, 24(2):290–292, 2008.