OXFORD

Sequence analysis

pycofitness—Evaluating the fitness landscape of RNA and protein sequences

Fabrizio Pucci [ID] [1,2,†], **Mehari B. Zerihun** [3,†], **Marianne Rooman** [1,2], **Alexander Schug** [ID] [3,4,*]

[1]Computational Biology and Bioinformatics, Université Libre de Bruxelles, 1050 Brussels, Belgium
[2]Interuniversity Institute of Bioinformatics in Brussels, 1050 Brussels, Belgium
[3]John von Neumann Institute for Computing, Jülich Supercomputer Centre, 52428 Jülich, Germany
[4]Department of Biology, University of Duisburg-Essen, D-45141 Essen, Germany

*Corresponding author. John von Neumann Institute for Computing, Jülich Supercomputer Centre, 52428 Jülich, Germany. E-mail: al.schug@fz-juelich.de (A.S)
†Equal contribution.
Associate Editor: Yann Ponty

Abstract

Motivation: The accurate prediction of how mutations change biophysical properties of proteins or RNA is a major goal in computational biology with tremendous impacts on protein design and genetic variant interpretation. Evolutionary approaches such as coevolution can help solving this issue.

Results: We present pycofitness, a standalone Python-based software package for the *in silico* mutagenesis of protein and RNA sequences. It is based on coevolution and, more specifically, on a popular inverse statistical approach, namely direct coupling analysis by pseudo-likelihood maximization. Its efficient implementation and user-friendly command line interface make it an easy-to-use tool even for researchers with no bioinformatics background. To illustrate its strengths, we present three applications in which pycofitness efficiently predicts the deleteriousness of genetic variants and the effect of mutations on protein fitness and thermodynamic stability.

Availability and implementation: https://github.com/KIT-MBS/pycofitness.

1 Introduction

Accurately predicting the impact of mutations on protein and RNA stability and function is a longstanding issue in computational biology. Achieving this goal would be beneficial for a huge series of applications such as protein design (Coluzza 2017, Pucci *et al.* 2022), interpretation of genetic variants (Gerasimavicius *et al.* 2020, Iqbal *et al.* 2020) and understanding antibiotic resistance (Woodford and Ellington 2007). The massive amount of sequences made available over the past decade through the advancement of next-generation sequencing technologies can be exploited to estimate the effect of mutations. Indeed, conserved sites in multiple sequence alignments (MSA) of homologous proteins or RNAs typically characterize functionally or structurally important sites. Therefore, mutations at these sites are not well tolerated and generally removed during evolution.

Coevolutionary approaches, in which correlations between pairs of sites in MSAs are considered, are an additional level of evolutionary information that can be extracted from sequence data. Direct coupling analysis (DCA) (Schug *et al.* 2009, Weigt *et al.* 2009, Morcos *et al.* 2011) is one of these methods. It uses an inverse statistical inference formalism to extract coevolutionary information and to identify pairs of MSA sites that coevolve.

Coevolutionary approaches and the information extracted from them have been extensively used as constraints in 3D structure modeling of proteins (Schug *et al.* 2009, Marks *et al.* 2011, Dago *et al.* 2012, Hopf *et al.* 2012, Morcos *et al.* 2013) and RNA (De Leonardis *et al.* 2015, Weinreb *et al.* 2016, Pucci and Schug 2019). They have also been employed in the computational characterization of the impact of mutations on protein fitness and biophysical characteristics and provide an efficient estimation of deep mutagenesis scanning data (Cheng *et al.* 2016, Figliuzzi *et al.* 2016, Flynn *et al.* 2017, Hopf *et al.* 2017, Peng *et al.* 2019, Frazer *et al.* 2021). Recently, they have been successfully used in the construction of generative protein sequence models for protein design applications (Russ *et al.* 2020, McGee *et al.* 2021, Trinquier *et al.* 2021).

To summarize, coevolutionary methods are of fundamental interest in a broad array of applications, including the understanding of genetic variant deleteriousness and, from a biotechnological perspective, the rational design of proteins and RNA.

2 Direct coupling analysis

DCA is a statistical method to extract coevolutionary information from an MSA of homologous RNA or protein sequences (Cocco *et al.* 2018). Let $P(S)$ be the probability that a given RNA (protein) sequence $S = a_1 a_2 \ldots a_L$ of length L, in which each state a_i is either an RNA base (an amino

acid residue) or a gap, is sampled over the course of evolution. $P(S)$ can be written using the Boltzmann law as:

$$P(S) = \frac{1}{Z}\exp(-\beta\phi), \qquad (1)$$

where β is the inverse of the temperature, Z the partition function of the model and ϕ the Hamiltonian of the system, which in turn is expressed via a generalized Potts model as:

$$\phi = -\sum_{i<j}^{L} J_{ij}(a_i, a_j) - \sum_{i=1}^{L} h_i(a_i). \qquad (2)$$

The parameters $h_i(a_i)$ measure the local field strength at site i occupied by state a_i and the coupling parameters $J_{ij}(a_i, a_j)$ quantify the coupling strength between pairs of sites i and j occupied by states a_i and a_j, respectively.

The local field and the coupling parameters of the model are inferred from the MSA of RNAs (proteins) that are homologous to S using inverse statistical algorithms such as message passing DCA (mpDCA) (Weigt *et al.* 2009), mean-field DCA (mfDCA) (Morcos *et al.* 2011), pseudo-likelihood maximization DCA (plmDCA) (Ekeberg *et al.* 2013), Boltzmann learning (Cuturello *et al.* 2020), Gaussian DCA (Baldassi *et al.* 2014) and autoregressive DCA (arDCA) (Trinquier *et al.* 2021).

Let us now consider a substitution at position i in the sequence S, with state a_i replaced by b_i. We can quantify the effect of the substitution on the evolutionary energy, $\Delta E(a_i \rightarrow b_i)$, by computing the change in ϕ between the wild type and mutated RNA (protein) as:

$$\Delta E(a_i \rightarrow b_i) = \phi(a_1 \ldots a_i \ldots a_L) - \phi(a_1 \ldots b_i \ldots a_L). \qquad (3)$$

With this convention, the lower the ΔE, the more deleterious the mutation is for the RNA (protein). Conversely, if ΔE is positive, the mutation is beneficial for the protein. In terms of the coupling and single-site field strengths, Equation (3) can be rewritten as:

$$\Delta E(a_i \rightarrow b_i) = h_i(b_i) - h_i(a_i) + \sum_{j=1, j\neq i}^{L} \left(J_{ij}(b_i, a_j) - J_{ij}(a_i, a_j) \right). \qquad (4)$$

This change in evolutionary energy is a measure of the fitness of the mutated RNA (protein) with respect to the wild type.

3 Implementation of pycofitness

We have developed pycofitness, a Python package to perform *in silico* mutagenesis experiments of protein or RNA sequences using a widely recognized DCA approach. Based on an input MSA of proteins or RNAs homologous to the target sequence, pycofitnesss calculates the couplings and single-site fields of a coevolutionary model employing the plmDCA algorithm that we have implemented earlier in the python package pyDCA (Zerihun *et al.* 2020). The strengths of plmDCA with respect to other DCA approaches are demonstrated in Supplementary Section S3. Details about the implementation and use of pycofitness are given in Supplementary Section S1; for technical details about the model and

parameter choice, such as the regularization terms, we refer the reader to Zerihun *et al.* (2020).

Once the model is inferred, pycofitness inserts, in turn, all possible single-site mutations in the reference sequence, i.e. three nucleotide substitutions per site for RNAs and 19 amino acid substitutions for proteins. pycofitness estimates the effect of a nucleotide or amino acid substitution as in Equation (4), thus by computing the difference in evolutionary energy between the wild-type and mutated sequences, using the inferred coevolutionary model.

The pycofitness software has several strengths and capabilities:

- First and foremost, it is a standalone Python software that can be easily installed from the Python packaging index (PyPI). It can be used through the command line or as a Python library. Its user interface can be effortlessly imported and integrated with user-developed Python source codes. Details are given in Supplementary Section S2.
- Moreover, the computationally demanding parts of the algorithm, such as parameter inference through plmDCA, are implemented using the C/C++ backend, enabling efficient code parallelization and allowing pycofitness to perform long protein and RNA mutagenesis in a reasonable amount of time.
- Another useful feature of pycofitness is its ability to perform RNA mutagenesis, a field of growing interest. Only a few tools are available for such studies, making pycofitness an invaluable addition to the current research landscape.
- pycofitness is also user-friendly: the only input it requires is the MSA and the type of biomolecule ("RNA" or "protein").

4 Applications of pycofitness

In addition to its ease of use, pycofitness has been extensively tested for its accuracy, as shown below and in Supplementary Section S2. Moreover, a novelty and advantage of pycofitness is that it can run coevolutionary mutagenesis analyses of both protein and RNA sequences, unlike similar packages such as EVcoupling (Hopf *et al.* 2018).

Here we applied pycofitness to non-coding RNA molecules, and in particular, tRNAs, which are essential molecules in the protein synthesis pathway and link the genetic information to the amino acid sequence of proteins. More specifically, we studied the effect of mutations on mitochondrially encoded tRNAs: tRNA-Phe (MT-TF), tRNA-Val (MT-TV), tRNA-Leu I (MT-TL1), and tRNA-Ile (MT-TI).

We started by collecting variants in these four genes from the ClinVar database (Landrum *et al.* 2018). We considered two classes of variants: the pathogenic class containing the variants assigned as pathogenic and likely pathogenic, and the benign class grouping likely benign and benign variants. Variants in these genes have been related to a wide variety of genetic diseases such as mitochondrial encephalomyopathy, lactic acidosis, stroke-like episodes (MELAS), myoclonic epilepsy with ragged-red fibers (MERRF), mitochondrial cardiomyopathy, and complex IV deficiency of the mitochondrial respiratory chain (Abbott *et al.* 2014).

We computed the change in coevolutionary energy ΔE of the variants and used it to predict their deleteriousness. We

started by aligning the tRNA sequences from the human reference genome GRCh38 with the RFAM family RF00005 (Kalvari *et al.* 2021) using the RNA homologous sequence inference model Infernal (Nawrocki and Eddy 2013). Once the MSA was constructed, we run pycofitness to compute the ΔE values. If ΔE is below a given threshold value, which was optimized for each gene in leave-one-out cross validation, the variants were considered as pathogenic; otherwise the variants were predicted as benign.

We used common metrics for binary classification to assess the accuracy of pycofitness, i.e. sensitivity [TP/(TP+FN)] and specificity [TN/(TN+FP)], where TP are the variants correctly predicted as pathogenic, TN, those correctly predicted as benign, FP, those wrongly classified as pathogenic and FN, those wrongly predicted as benign. We also computed the balanced accuracy (BACC) as the average between sensitivity and specificity, as well as the threshold-independent metric that is the area under the receiver operating characteristic curve (AUC-ROC).

As shown in Table 1, pycofitness is able to predict the deleteriousness of the variants in the tested non-coding RNAs with good accuracy, with BACC scores between 0.63 and 0.79 and AUC-ROC between 0.71 and 0.83. Moreover, we observe in Fig. 1 that the ΔE probability distributions of

pathogenic and benign variants are well separated for the four tRNAs.

We showcase two other applications of pycofitness in Supplementary Section S2. In the first, we used pycofitness to predict the impact of mutations on the fitness of four enzymes on which high-throughput fitness experiments have been performed (Weile *et al.* 2017), and compared its performance with that of six fitness predictors. We found that pycofitness achieves on average the same accuracy as the other unsupervised methods tested and, unsurprisingly, performs equally or slightly worse than the supervised methods, which are based on complex machine learning models and trained on huge amounts of variant information.

In the last application, we used pycofitness to predict the change in folding free energy ($\Delta\Delta G$) upon all possible single-site mutations inserted in the β_1 domain of streptococcal protein G, and compared them with $\Delta\Delta G$s measured by high-throughput experiments (Nisthal *et al.* 2019). We found that pycofitness scores correlate less well with experimental $\Delta\Delta G$s than $\Delta\Delta G$s obtained with the best supervised structure-based predictors. However, we showed how we can integrate the pycofitness scores into these other predictors to boost the overall performances.

In summary, we showed that pycofitness performs very well in protein and RNA fitness and deleteriousness predictions. For protein $\Delta\Delta G$ predictions, it performs less well than supervised structure-based methods dedicated to $\Delta\Delta G$ predictions. This is not surprising as pycofitness is based on evolutionary information and thus predicts fitness rather than stability, which are only partially correlated. For example, functional residues are highly conserved but not at all optimized for stability, usually being stability weaknesses of the protein structure (Ferreiro *et al.* 2018, Hou *et al.* 2021). However, pycofitness scores and predicted $\Delta\Delta G$s are complementary and their combination yield increased predictions accuracy.

Table 1. Sensitivity, specificity, balanced accuracy, and area under the receiver operating characteristic curve of pycofitness in predicting the deleteriousness of variants with clinical annotations in non-coding RNAs.

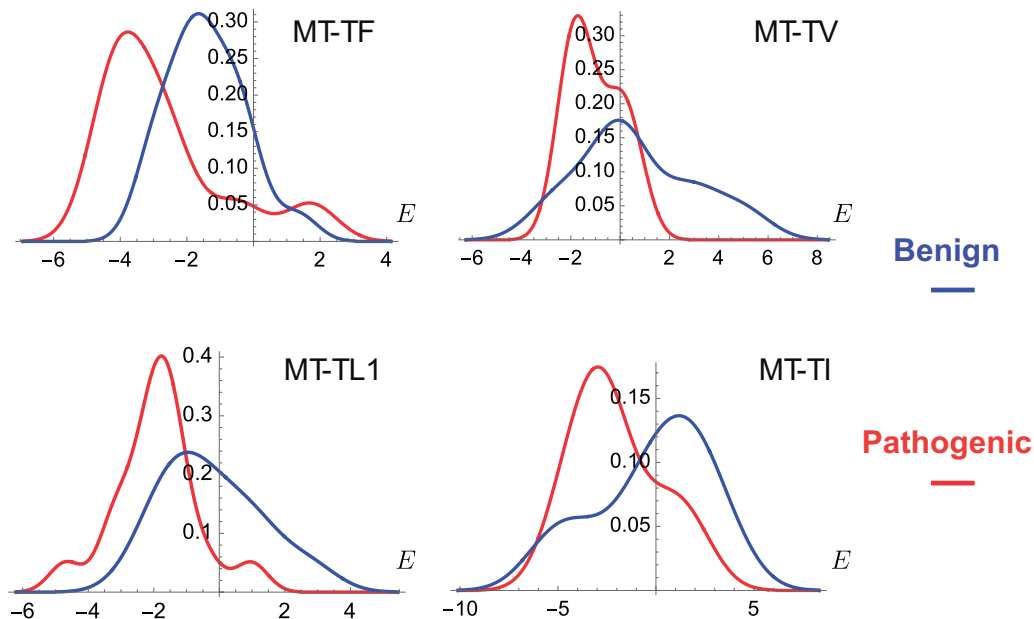| Type	Sensitivity	Specificity	BACC	AUC-ROC	No. of variants
MT-TF	0.80	0.79	0.79	0.77	29
MT-TV	0.60	0.67	0.63	0.71	14
MT-TL1	0.69	0.80	0.74	0.83	26
MT-TI	0.71	0.66	0.69	0.70	16

Figure 1. Probability distribution of the change in coevolutionary energy ΔE for pathogenic and benign variants in four mitochondrially encoded tRNAs: tRNA-Phe (MT-TF), tRNA-Val (MT-TV), tRNA-Leu I (MT-TL1), and tRNA-Ile (MT-TI).

Detailed results of these analyses are available in our GitHub repository at https://github.com/KIT-MBS/pycofitness.

5 Conclusion

In this applications note, we have introduced pycofitness, a Python package for in-silico mutagenesis of RNA and protein sequences which uses an MSA of homologous sequences as input, and made it available to the scientific community. Its user-friendly installation, ease of use, and good performances make it very valuable, by itself or in combination with other tools, in a wide range of applications ranging from the interpretation of genetic variants to the rational modification of proteins and RNAs. We would like to emphasize that pycofitness is unsupervised and thus does not suffer from bias issues.

Acknowledgements

Supplementary data

Supplementary data are available at *Bioinformatics* online.

Conflict of interest

None declared.

Funding

Data availability

The data underlying this article are available in the article, in its online supplementary material and in the repository https://github.com/KIT-MBS/pycofitness.

References

Abbott JA, Francklyn CS, Robey-Bond SM. Transfer RNA and human disease. *Front Genet* 2014;**5**:158.

Baldassi C, Zamparo M, Feinauer C *et al.* Fast and accurate multivariate gaussian modeling of protein families: predicting residue contacts and protein-interaction partners. *PLoS One* 2014;**9**:e92721.

Cheng RR, Nordesjö O, Hayes RL *et al.* Connecting the sequence-space of bacterial signaling proteins to phenotypes using coevolutionary landscapes. *Mol Biol Evol* 2016;**33**:3054–64.

Cocco S, Feinauer C, Figliuzzi M *et al.* Inverse statistical physics of protein sequences: a key issues review. *Rep Prog Phys* 2018;**81**:032601.

Coluzza I. Computational protein design: a review. *J Phys Condensed Matter* 2017;**29**:143001.

Cuturello F, Tiana G, Bussi G. Assessing the accuracy of direct-coupling analysis for RNA contact prediction. *RNA* 2020;**26**:637–47.

Dago AE, Schug A, Procaccini A *et al.* Structural basis of histidine kinase autophosphorylation deduced by integrating genomics, molecular dynamics, and mutagenesis. *Proc Natl Acad Sci USA* 2012;**109**:E1733–42.

De Leonardis E, Lutz B, Ratz S *et al.* Direct-coupling analysis of nucleotide coevolution facilitates RNA secondary and tertiary structure prediction. *Nucl. Acids Res* 2015;**43**:10444–55.

Ekeberg M, Lövkvist C, Lan Y *et al.* Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Phys Rev E* 2013;**87**:1–16.

Ferreiro DU, Komives EA, Wolynes PG. Frustration, function and folding. *Curr Opin Struct Biol* 2018;**48**:68–73.

Figliuzzi M, Jacquier H, Schug A *et al.* Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase TEM-1. *Mol Biol Evol* 2016;**33**:268–80.

Flynn WF, Haldane A, Torbett BE *et al.* Inference of epistatic effects leading to entrenchment and drug resistance in HIV-1 protease. *Mol Biol Evol* 2017;**34**:1291–306.

Frazer J, Notin P, Dias M *et al.* Disease variant prediction with deep generative models of evolutionary data. *Nature* 2021;**599**:91–5.

Gerasimavicius L, Liu X, Marsh JA. Identification of pathogenic missense mutations using protein stability predictors. *Sci Rep* 2020;**10**:15387–10.

Hopf TA, Colwell LJ, Sheridan R *et al.* Three-dimensional structures of membrane proteins from genomic sequencing. *Cell* 2012;**149**:1607–21.

Hopf TA, Ingraham JB, Poelwijk FJ *et al.* Mutation effects predicted from sequence co-variation. *Nat Biotechnol* 2017;**35**:128–35.

Hopf TA, Green AG, Schubert B *et al.* The EVcouplings Python framework for coevolutionary sequence analysis. *Bioinformatics* 2018;**35**:1582–4.

Hou Q, Pucci F, Ancien F *et al.* SWOTein: a structure-based approach to predict stability strengths and weaknesses of proteins. *Bioinformatics* 2021;**37**:1963–1971.

Iqbal S, Pérez-Palma E, Jespersen JB *et al.* Comprehensive characterization of amino acid positions in protein structures reveals molecular effect of missense variants. *Proc Natl Acad Sci USA* 2020;**117**:28201–11.

Kalvari I, Nawrocki EP, Ontiveros-Palacios N *et al.* Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res* 2021;**49**:D192–D200.

Landrum MJ, Lee JM, Benson M *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* 2018;**46**:D1062–7.

Marks DS, Colwell LJ, Sheridan R *et al.* Protein 3D structure computed from evolutionary sequence variation. *PLoS One* 2011;**6**:e28766.

McGee F, Hauri S, Novinger Q *et al.* The generative capacity of probabilistic protein sequence models. *Nat Commun* 2021;**12**:6302.

Morcos F, Pagnani A, Lunt B *et al.* Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci USA* 2011;**108**:E1293–301.

Morcos F, Jana B, Hwa T *et al.* Coevolutionary signals across protein lineages help capture multiple protein conformations. *Proc Natl Acad Sci USA* 2013;**110**:20533–8.

Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 2013;**29**:2933–5.

Nisthal A, Wang CY, Ary ML *et al.* Protein stability engineering insights revealed by domain-wide comprehensive mutagenesis. *Proc Natl Acad Sci USA* 2019;**116**:16367–77.

Peng M, Maier M, Esch J *et al.* Direct coupling analysis improves the identification of beneficial amino acid mutations for the functional thermostabilization of a delicate decarboxylase. *Biol Chem* 2019;**400**:1519–27.

Pucci F, Schug A. Shedding light on the dark matter of the biomolecular structural universe: progress in RNA 3D structure prediction. *Methods* 2019;**162-163**:68–73.

Pucci F, Schwersensky M, Rooman M. Artificial intelligence challenges for predicting the impact of mutations on protein stability. *Curr Opin Struct Biol* 2022;**72**:161–8.

Russ WP, Figliuzzi M, Stocker C *et al.* An evolution-based model for designing chorismate mutase enzymes. *Science* 2020;**369**:440–5.

Schug A, Weigt M, Onuchic JN *et al.* High-resolution protein complexes from integrating genomic information with molecular simulation. *Proc Natl Acad Sci USA* 2009;**106**:22124–9.

Trinquier J, Uguzzoni G, Pagnani A *et al.* Efficient generative modeling of protein sequences using simple autoregressive models. *Nat Commun* 2021;**12**:5800.

Weigt M, White R. A, Szurmant H *et al.* Identification of direct residue contacts in protein–protein interaction by message passing. *Proc Natl Acad Sci USA* 2009;**106**:67–72.

Weile J, Sun S, Cote AG *et al.* A framework for exhaustively mapping functional missense variants. *Mol Syst Biol* 2017;**13**:957.

Weinreb C, Riesselman AJ, Ingraham JB *et al.* 3D RNA and functional interactions from evolutionary couplings. *Cell* 2016;**165**:963–75.

Woodford N, Ellington MJ. The emergence of antibiotic resistance by mutation. *Clin Microbiol Infect* 2007;**13**:5–18.

Zerihun MB, Pucci F, Peter EK *et al.* pydca v1.0: a comprehensive software for direct coupling analysis of RNA and protein sequences. *Bioinformatics* 2020;**36**:2264–5.