Generate new synthetic protein with Machine Learning

Abstract

Authors

Capucine GRANCLAUDE
Chourouk JDIDI
Gül Vahide YILMAZ
Loïc LECOMTE
Baudouin DE LAVIGNE

Master 2 students, in the field of Bio-informatic at the engineering school EFREI PARIS, Paris, FRANCE.

Date

September 2022 - February 2023

1 Introduction

"What I cannot create, I do not understand", we can summarize the principle of synthetic biology by this quote from Richard Feynman.[1] Indeed, we can explain synthetic biology as the design and construction of new biological systems, and the reassembly of those naturally existing for the needs of useful applications from known elements. We can compare it to Lego bricks that we assemble to create a universe. [2]

This term encompasses many methods using knowledge from biology, computer science and engineering. Synthetic biology has a major influence on various sectors by creating the conditions for the efficient production of products that are essential to society, such as new therapeutic substances or fuels. It also contributes to a better understanding of fundamental life processes.

Machine Learning is a scientific field that consists in letting algorithms discover patterns, recurring motifs, from datasets. This data can be numbers, words, images, statistics...

With the resolution of patterns in these data, the algorithms learn and improve their performance autonomously in the execution of a specific task.
Finally, Machine Learning algorithms learn to perform a task or make predictions from data and improve their performance over time.

Our main objective during this project is to generate new protein being derivate from existing protein. For that we will use generative model that are able to identify important part of a protein to keep them and then add variation to create a new protein. We will try to create new luciferase-like oxidoreductases which is a protein taht have a measurable luminescence. This luminescence will be use to estimate if the generated protein is still functional.

2 State of art

Synthetic biology

Today, synthetic biology allows us to build a genome from A to Z by choosing the desired genes and their order. To achieve this, there are several ways with more or less efficient methods that we will describe.

First of all, the pre-logical way consists in using the vocation of the cells to reproduce in two cells. Indeed, we will select the desired DNA fragment and then introduce it into a bacterium. The bacteria are allowed to multiply to a large number (1 billion for example) and then the numerous fragments obtained are extracted from the bacteria. This method is however long and requires many manipulations. [3]

Secondly, PCR (polymerase chain reaction) is a technique that allows the reproduction of DNA sequences in large quantities. Democratized since the use of PCR test for Covid-19, it allows the creation of numerous copies of DNA in only three steps. This method is widely used by laboratories for the detection of infectious diseases because it simplifies the study of a DNA sample. However, it remains only a copy of a DNA fragment. Moreover, the error rate is high with the pollution of the amplicons. [4]

Oligonucleotides are single-stranded DNA fragments of a size of about ten nucleotides, which is small. They are obtained by chemical synthesis of a single strand from functional groups chosen for their interest. They are used for example as a primer in PCR or in DNA chips. [5],[6]

DNA chips are programmable chips to synthesize oligonucleotides, amplify them, select them by hybridization. This allows them to be assembled and reduces the error rate which is quite high for simple oligonucleotides. [7] However, oligonucleotides are very expensive for laboratories and the error rate remains high. This is why new technologies allow synthetic biology to reduce human manipulation errors and make it more accessible.

Concerning protein creation in the field of synthetic biology, there is a field called protein engineering.

This field allows to participate in the understanding of protein folding and the recognition of protein design principles. It is a booming sector with an estimated $168 billion in 2017. [8]

Two strategies exist for protein engineering: rational protein design and directed evolution. They are generally used in parallel by researchers using both 2D and 3D structure.
Proteins can be designed from scratch or by making computed variants of a known protein structure and its sequence. Rational design allows the prediction of the sequence of proteins that fold into specific structures. The prediction of these sequences can then be validated experimentally by methods such as directed mutagenesis, peptide synthesis or artificial gene synthesis.

This method is therefore performed manually in the laboratory through long protein studies. This work is therefore long and meticulous. This is why today researchers are looking to automate this process in order to accelerate and improve the results of these methods.

Machine learning

We will quickly explain what machine learning is before discussing applications to synthetic biology. Machine learning is a computer programming technique that uses statistical probabilities to give computers the ability to learn on their own without explicit programming. The basic goal of machine learning is to “teach computers how to learn” – and then to act and react – as humans do, improving their learning style and knowledge autonomously over time.

As seen earlier, synthetic biology is a rapidly growing field that uses genetic engineering techniques to design and manufacture artificial molecules, cells and living organisms.
Machine learning is increasingly used in this field to solve problems such as protein design, drug discovery and modeling of biological systems.

Supervised machine learning is an elementary but strict technology. Operators present the computer with sample inputs and the desired outputs, and the computer searches for solutions to get those outputs based on those inputs. The goal is for the computer to learn the general rule that maps inputs and outputs.
Supervised machine learning can be used to make predictions about unavailable or future data (this is called "predictive modeling"). The algorithm tries to develop a function that accurately predicts the output from the input variables. [9]

Supervised machine learning can be divided into two types (Fig1) :

- Classification: the output variable is a category.
- Regression: the output variable is a specific value.

The main algorithms of supervised machine learning are: random forests, decision trees, K-NN (k-Nearest Neighbors) algorithm, linear regression, Naïve Bayes algorithm, support vector machine (SVM), logistic regression and boosting gradient.

There are several machine learning approaches that are used in synthetic biology, such as neural networks, decision trees, genetic algorithms, and regression systems.
There are many challenges in synthetic biology that can be solved using machine learning techniques. For example, the optimization of protein production using machine learning algorithms, pattern recognition for the detection of disease-associated genes, and the modeling of biological systems for understanding biology at a systems level.
In general, machine learning is becoming an increasingly important tool for synthetic biology, allowing complex problems to be solved and production processes to be optimized.
Synthetic biologists are beginning to take advantage of deep learning methods, fueled by advances in synthesis and sequencing.

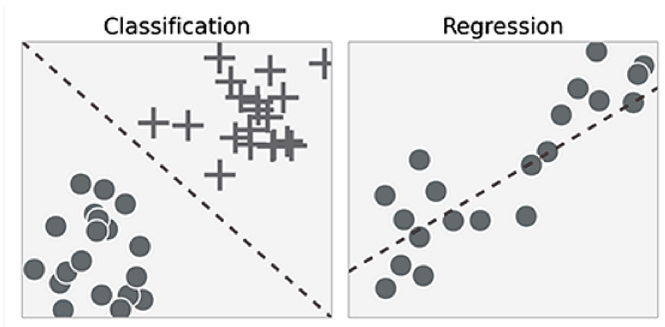

Fig 1 *- Illustration of difference between linear classification and linear regression*

Deep Learning

Deep learning is one of the main technologies of machine learning. With Deep Learning, we are talking about algorithms capable of mimicking the actions of the human brain through artificial neural networks (***Fig 2***). Networks are made up of tens or even hundreds of "layers" of neurons, each receiving and interpreting information from the previous layer.

Each artificial neuron can be seen as a linear model (***Fig 3***). By interconnecting the neurons as a layer, we transform our neural network into a very complex non-linear model.

Deep learning models tend to perform well with large amounts of data, whereas more traditional machine learning models stop improving after a saturation point. [10]

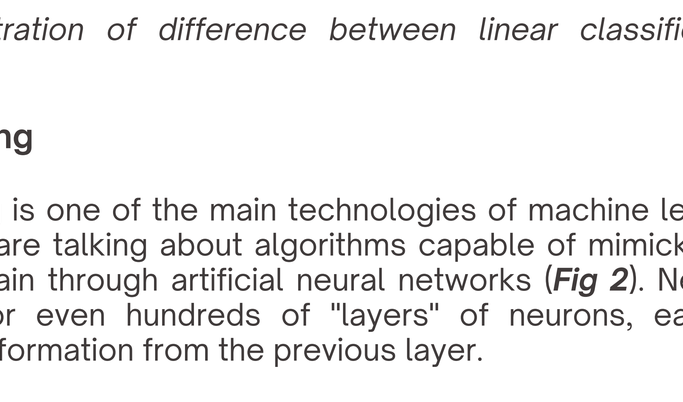

Fig 2 *- Neural Network architecture*

Today, the mastery of these two fields can be an issue because it is necessary to be up to date with both fields, synthetic biology and Deep Learning.
Deep Learning and Machine Learning need a lot of data which can be an issue in some synthetic biology case.

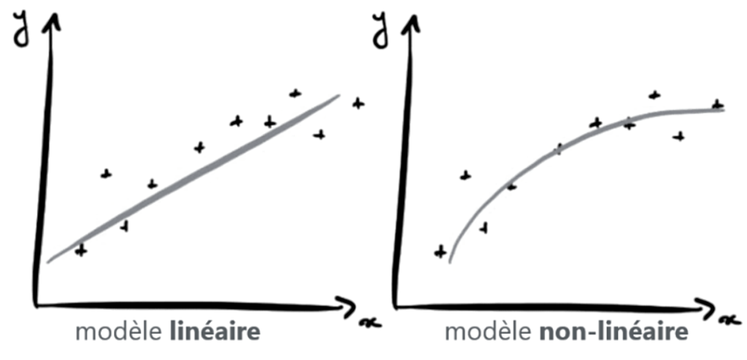

Fig 3 *- Linear model and non-linear model [9]*

Deep Learning around synthetic biology

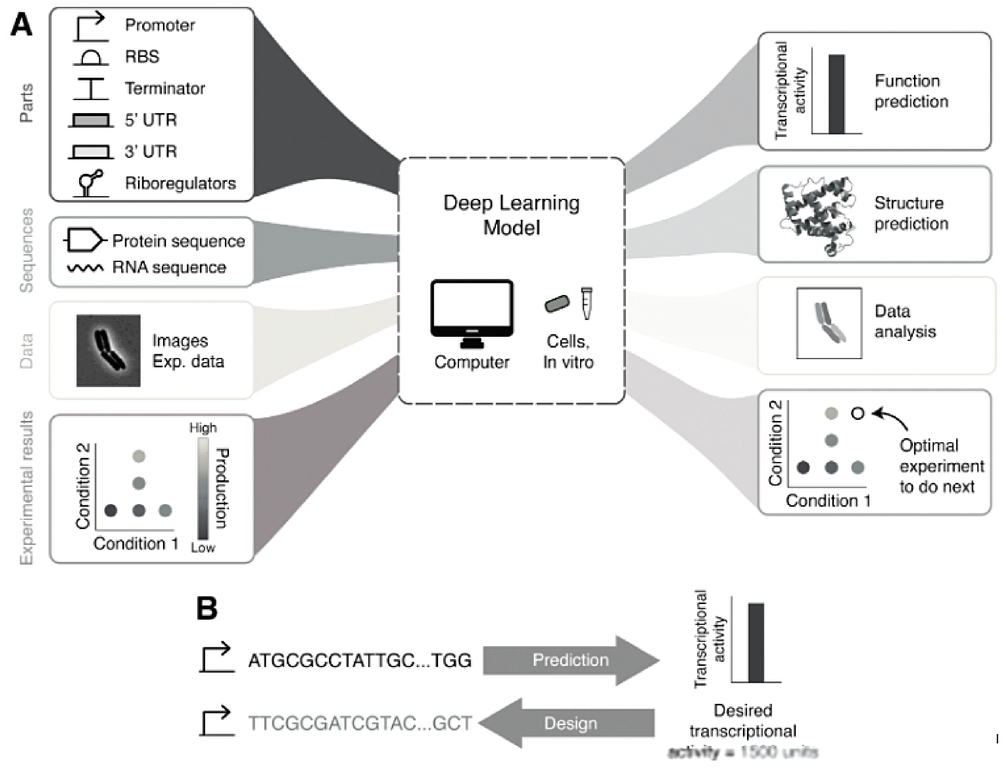

Fig 4 *- Deep learning applications in synthetic biology. [10]*

We can see examples of relevant inputs for deep learning networks and their associated output predictions. (***Fig 4***)
For example, deep learning can be used to predict the function of biological “parts” such as promoters, ribosome binding sites or even untranslated regions. [11]
Sample and all developed a model that accurately predicts how the 5' UTR sequence controls ribosomal loading.

For synthetic biology applications, it is often desirable that a model be not only predictive (eg, from sequence to predicted performance), but also generative. Non-deep learning examples have proven to be invaluable to the biological engineering community. For example, the RBS 55 calculator can generate new designs based on a thermodynamic model, and synthetic 5' UTR sequences have been successfully made based on genetic algorithms. [12]

Mechanistic modeling approaches are very powerful; however, they require in-depth knowledge of features that contribute to performance. Generative approaches based on deep learning models are an exciting area of development, as these tools evolve the ability to work backwards, from translation efficiency specifications to candidate sequence designs. For example, in their study of yeast promoters, Kotopka and Smolke used a CNN model to implement sequence design strategies, ultimately showing that the best algorithms produced strong constitutive and inducible synthetic promoters. [13]

Rapid advances in the field of geometric deep learning have facilitated an explosion of structure-function learning research in biotechnology. Perhaps the most publicized case is that of DeepMind's AlphaFold protein structure prediction model, which offers high enough protein structure prediction accuracy to be usable as a replacement for protein crystallography which is costly and time-consuming. The model takes protein sequence and multiple sequence alignments to similar proteins as inputs, and trains on three different data structures: a sequence-level representation, a nucleotide interaction representation by pairs and the 3D structure at the atom level of the protein generated by the model .

Although deep learning models are typically implemented using computers, several recent studies have demonstrated that ANN mimics can be constructed using biomolecular components. These models design biochemical systems and living cells that can perform calculations and “learn” to solve simple benchmark optimization problems. One of the main reasons this is possible is because inducible gene responses to chemical inducers typically resemble a sigmoid function of inducer concentration, and therefore can serve as a nonlinear function in the neural model. [14]

Based on this, Moorman presented the theoretical architecture of a biomolecular neural network, a dynamic chemical reaction network that faithfully implements ANN calculations, and demonstrated its use for the tasks classification. The authors highlighted the usefulness of molecular sequestration to achieve negative weight values and sigmoidal activation function in its elementary unit called biomolecular perceptron. Following this, Samaniego theoretically demonstrated that interconnected phosphorylation/dephosphorylation cycles can function as multi-layered biomolecular neural networks.

Generative models

Generative models are model able to extract characteristics from a dataset and then create new data presenting these characteristics and some variations. These new data are then similar to our original dataset but present some variations so they are new data. [15]

We can find these kind of generative model on this website [16] where we create a face of a person that does not exist, that face does not come from a database containing photographs of faces taken from an actual person. At each connexion we can create a different face.

We used this kind of model to create new proteins that will keep the principal functions of a family of proteins but presenting new functions or less secondary effects.

We have identified 2 models able to generate new datas the GANs and the VAEs.

GANs : generative adversarial network

GANs model are in reality constituted of two neural network. The first one called generator, is responsible for creating the new data. The second called discriminator is responsible for identifying if a data is real or invented.

When we train a GANs our purpose is that the generator is able to create data so close to real data that our discriminator identified them as real data.

To train a GANs model we first train the discriminator by providing it the real data so it learn to recognize real data. Then we ask the generator to create random data and we ask the discriminator to classify it as real or invented data. Then the discriminator give a feedback to the generator. With this feedback the generator try to create new data that are better than the first batch of invented data. We will repeat this operation until the invented data are good enough and are able to fool the discriminator. When the discriminator is not able to classify invented data as invented data then it mean that our data are closed enough from reality. [17] (***Fig 5***)

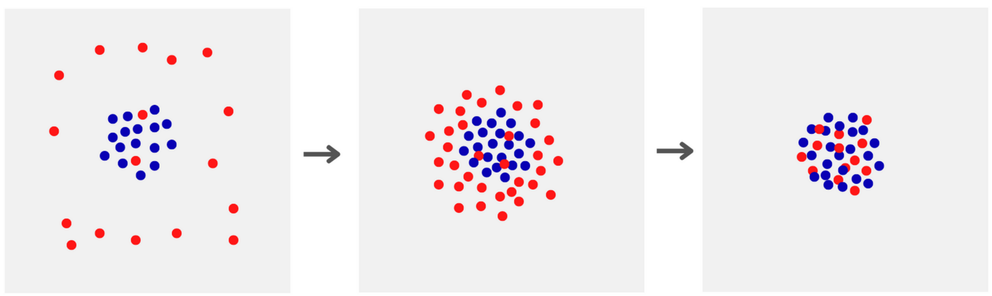

Fig 5 : *visualization of invented data during GANs's training*
In blue the real data and in red the invented data. Left is the beginning of GAN's training and right the end of the training.

VAE : Variational Autoencoders

VAE are a derived from Autoencoders (AE), a neural network that can be used in other fields than data creation. AE model are also constitued of two neural network the encoder and the decoder.

Encoder mission is to reduce the number of features used to describe the data without losing information. We passed from a data's description with X feature to a description with Y features with X > Y.
Decoder's mission is the reverse we passed from a data's description with Y features to a description with X features with Y < X.

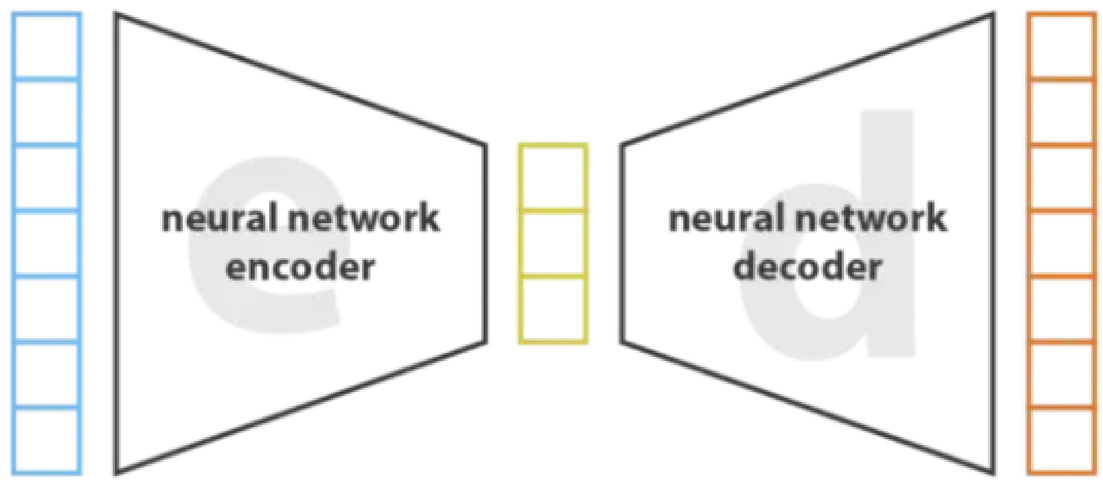

Fig 6 : *Autoencoder architecture [18]*

When we train an AE model the goal is to have for input a data with X features (***Fig 6*** : in blue) then reduce the feature's number to Y features (***Fig 6 :*** *in yellow*) using the encoder. Then with the decoder we try to recreate the exact same data with X features. The objectif is to find a representation with Y features without losing any information compared to the X feature représentation.

To create new data using Autoencoder we will first train the Autoencoder using the dataset of real data that we have. When the training is over and our AE is able to reduce the feature's number (encoder) and then expand the feature's number to recreate the original data (decoder), we keep only the decoder part. Then we give to the decoder random data with Y feature and then the decoder will transform it to a data in X feature similar to the real data.

But the Autoencoder have an issue when creating data which is that we can create aberrant data that are not similar to the real data. These aberrant data are due to the random number that we provide to the decoder. If these number are too different from the real data the decoder will not be able to transform it in a data similar to the real data. That's why Variational Autoencoders have been created. VAE model add a regularisation step between the encoder and the decoder. (***Fig 7***)

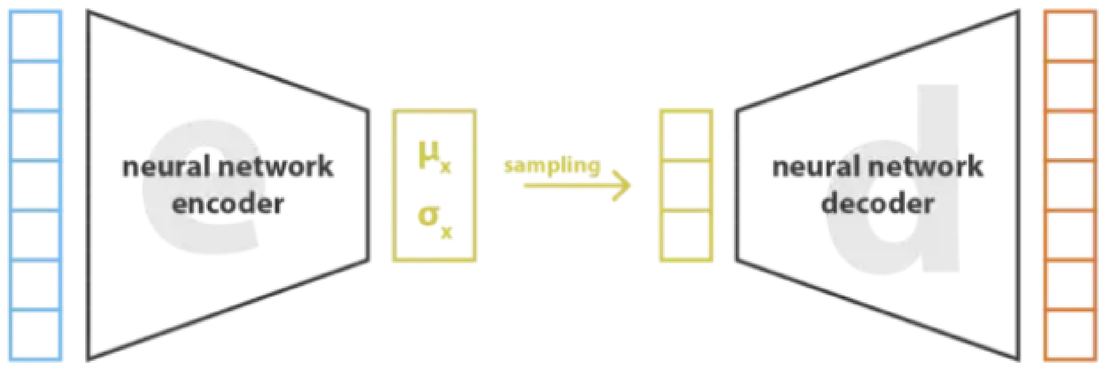

Fig 7 : *Variational Autoencoder architecture [18]*

This normalisation allow us to greatly reduce the possibility of aberrant data because even if we give to the decoder random number that are not similar to the real data the regularisation step will transform these number to number closer to the real data and then limit the creation of aberrant data. (***Fig 8***)

Fig 8 *: Exemple of data without regularization (left) and with regularization(right) If the created data is in the red, blue or yellow color a good data is created il it's in the white an aberrant data is created. We can see that with the regularization the probability of aberrant data is reduced. [18]*

7 Sources

Sources

[1] https://amphisciences.ouest-france.fr/2019/08/13/quest-ce-que-la-biologie-synthetique/

[2] https://sciencesnaturelles.ch/synthetic-biology-explained/what_is_synthetic_biology_

[3] https://www.sciencesetavenir.fr/videos/comment-fait-on-pour-synthetiser-de-ladn_8kxlxs

[4] https://fr.orbiere.com/pcr-reaction-en-chaine-de-la-polymerase/

[5] https://www.futura-sciences.com/sante/definitions/genetique-oligonucleotide-215/

[6] https://www.futura-sciences.com/sante/actualites/genetique-synthese-adn-laboratoire-passe-vitesse-superieure-5299/

[7] https://www.futura-sciences.com/sante/actualites/medecine-puce-adn-pourra-t-elle-jour-etablir-portrait-robot-43997/

[8] https://www.genengnews.com/magazine/239/speeding-up-the-protein-assembly-line/

[9] https://datascientest.com/machine-learning-tout-savoir?fbclid=IwAR39efgg4zWfgiOT2qbE9jCDNixHzpNBxvvM9ipt8wKFNTO7a7_eAz9PMVU

[10] https://datascientest.com/deep-learning-definition?fbclid=IwAR3te-V2cTKvrnxJ4_u3Kc8xWd2DclWK7c1gXvAahXyrwLI6oPU5NFLvIiQ

[11] https://www.liebertpub.com/doi/10.1089/genbio.2022.0017

[12] https://www.nature.com/articles/s41587-019-0164-5

[13] https://www.nature.com/articles/s41467-020-15977-4

[14] https://www.nature.com/articles/s41477-021-00932-y?fbclid=IwAR2SvP0jg2ueI5k41kRkd6-AJLCkbXGKOHkSF4P99smsN19qVpWIZp6suNg

[15] https://datascientest.com/generative-model-ou-modele-generatif

[16] https://thispersondoesnotexist.com/

[17] https://towardsdatascience.com/understanding-generative-adversarial-networks-gans-cd6e4651a29

[18] https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73