# Bioinformatics for Aspiring Synthetic Biologists

*Edgar Andrés Ochoa Cruz, Sayane Shome, Pablo Cárdenas, Maaruthy Yelleswarapu, Jitendra Kumar Gupta, Eugenio Maria Battaglia, Alioune Ngom, Pedro L. Fernandes, and Gerd Moe-Behrens*

## Abstract

For a synthetic biologist or biohacker to be able to hack, design, create, and engineer biological systems, the ability to work with biological data is essential. Basic bioinformatics skills will be required in order to read, interpret, write, and generate files containing DNA, RNA, protein, and other biological information. In this article, we will show the path you need to follow to implement a biological function using online data. As a case study, we are using Imperial College's 2014 iGEM project, which focused on the optimization of bacterial cellulose production for use in water filtration.

## Introduction

There are three requirements for the design of a biological system:

- The specification of the desired system in terms of its functions, inputs, and outputs
- The use of bioinformatics skills to select and combine DNA parts that follow these specifications
- The actual genetic modification of the organism in the wet lab

The design requires more than knowing how to pick bioparts from a catalog; it implies knowing how to create them and how to combine them to achieve the desired system.

Proper bioinformatics skills allow you to extract information within biological data and use it to model the desired system. Therefore, they're of top importance for any good biohacker.

There are large repositories of bioinformatics tools for synthetic biologists. Our main goal here is to describe the pipeline (see Figure 2-1) that allows a biohacker to use some of these tools in order to complete a synthetic biology project. We will focus on a practical example: the Imperial College's 2014 iGEM project. A video presentation of this project can be found on YouTube.

The Imperial College team focused on the biosynthesis and optimization of bacterial cellulose production using *Escherichia coli*. This biomaterial is used in industry for several purposes. Compared with plant cellulose, bacterial cellulose has advantageous properties, such as its high purity and strength, and its special porosity, which is what interested Imperial's team in the material as a potential water filter. The group was interested in the porosity characteristics that could make it useful as water-filter material to address a worldwide issue, water contamination. With that goal, they tried to reduce the cost of the bacterial cellulose production, which is a main limitation for commercialization of traditional cellulose filters.

In this article, we provide a general guide to performing fundamental *in silico* tasks for the development of one of the bioparts needed for this synthetic biology application. A more detailed guide with command-line details can be found on the Leukippos Institute home page.

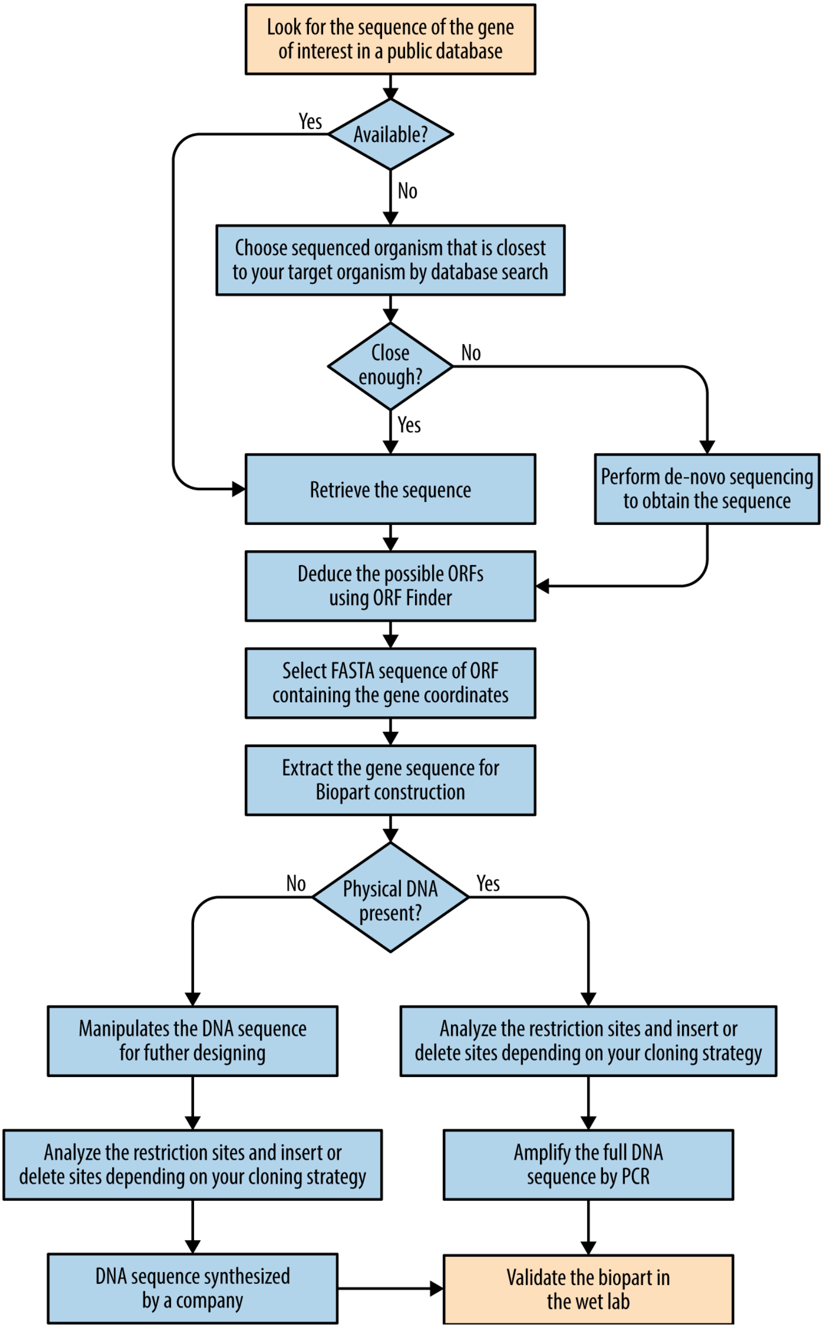

Look for the sequence of the gene of interest in a public database
Available?
Yes
No
Choose sequenced organism that is closest to your target organism by database search
Close enough?
No
Yes
Retrieve the sequence
Perform de-novo sequencing to obtain the sequence
Deduce the possible ORFs using ORF Finder
Select FASTA sequence of ORF containing the gene coordinates
Extract the gene sequence for Biopart construction
Physical DNA present?
No
Yes
Manipulates the DNA sequence for futher designing
Analyze the restriction sites and insert or delete sites depending on your cloning strategy
Analyze the restriction sites and insert or delete sites depending on your cloning strategy
Amplify the full DNA sequence by PCR
DNA sequence synthesized by a company
Validate the biopart in the wet lab

*Figure 2-1. Pipeline for executing a synthetic biology project applying bioinformatic tools*

## From Data to Life

For this purpose, bioinformatics tools can be divided in two main groups. The first group comprises analytical tools, which help you to understand the properties of biological data. Among these tools, you can find:

- DNA and RNA folding: the visualization of predicted secondary structures or hybridization patterns formed by DNA or RNA sequences is crucial to avoid the design being ruined by the presence of unforeseen structural aberrations. These could lead to a variety of disastrous results, from unsuccessful polymerase chain reactions (PCRs) to misfolded proteins. The University of Albany's RNA Institute website is a good place to start.
- Protein modeling: protein structure plays an important role in determining the biological functions of this biopart. Theoretical models are predicted for amino acid sequences for which experimentally determined structures are not known yet. Various techniques are employed for protein modeling, and they differ in complexity, accuracy, and the primary data required for the structure prediction. We recommend that you start with these:
  - Homology modeling: SWISS-MODEL is an automated protein structure homology modeling server that accepts requests either through the Expasy server or can be accessed via the eepView program. To perform protein modeling, the target amino acid sequence and 3D structure of a template protein are required. To determine the template protein structure, perform a BLAST (tool for searching online databases, based on nucleotide or amino acid homology) search with the target sequence against the database of previously determined protein structures (.pdb). Use the structure with maximum identity (homology) score as the template for the modeling.
  - Validation of the modeled protein structure: validate the modeled structure via the SAVES server. The server validates the model based on six parameters and suggests if the model satisfies the essential requirements to be a good theoretical model.

The second group of tools contains the software that helps you to manipulate sequence data in order to create and design new bioparts (genes, promoters, ribosome binding sites, etc.). One could say that these are the tools that actually help you to engineer life. In this guide, we will focus on these tools to show you how to produce a biopart in your hackerspace.

Most of genetic information you need is already available online and is free to access. Among the databases that you can use, the most popular ones are:

- GENBANK, a database of gene sequences hosted by the U.S. National Center for Biotechnology Information (NCBI).
- UniPROT, a database that provides a clustered sets of sequences from the UniProtKB and selected UniProt Archive records to obtain complete coverage while hiding redundant sequences.
- ENSEMBL or UCSC, the main genome browsers with integrated knowledge bases. Though you can find species-specific genomic databases with a quick Google search, depending on the organism of interest, you'll find them much more reliably here.

For synthetic biology, we suggest the Registry of Biological Parts and the Joint BioEnergy Institute's registry (JBEIR).

In some cases, the information is not available online, so you will need to have access to the physical DNA or organism from which you can extract the genomic DNA. Besides genomic DNA extraction, in certain contexts (e.g., gene cloning, gene probes, or creation of cDNA libraries) you might want to synthesize complementary DNA (cDNA). The cDNA is double-stranded DNA synthesized from mature messenger RNA (mRNA) by using reverse transcriptase.

Nevertheless, you will eventually need to find the sequenced organism closest to your target organism using a database search and use its sequence to design primers to amplify your biopart from the extracted DNA using PCR. The following tools can be used to design your primers, using a homologous region containing your biopart, shared by sequenced and unsequenced organisms, as a virtual template: Primer3 and Primer–BLAST.

The biopart you want to amplify will probably not be identical in both organisms; therefore, the design of degenerate primers (defined as: "a population of specific primers that cover all the possible combinations of nucleotide sequences coding for a given protein sequence" by Iserte et al., 2013) is recommended.

In our practical example, Imperial's team wanted to amplify parts from the *Gluconacetobacter* genome and introduce them into the *E. coli*. *Gluconacetobacter* is characterized by its high production of bacterial cellulose but is more difficult to manipulate and scale in a biofactory than *E. coli* because of the lack of specific genetic tools and information. Therefore, the group decided to isolate and sequence a new strain, which they named *G. xylinus igem*, as well as strain ATCC 53852 in order to obtain new information. The group sequenced both bacteria to obtain the genetic information that allowed them to access the data. They also created several tools to genetically manipulate *Gluconacetobacter*. Sequencing is the third possibility for retrieving biological data, which is useful in case you don't find the data online for your organism or a closely related one. By the way, sequencing the gene of interest is always good practice to make sure you are working with what you think you are.

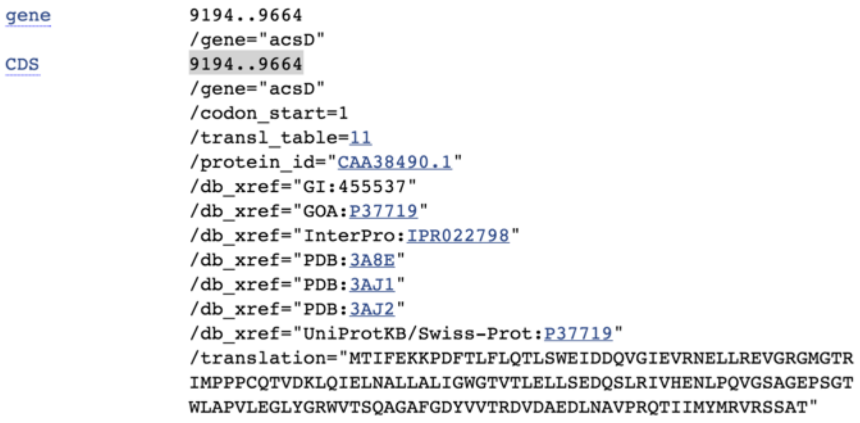
## One Biopart at a Time

The iGEM team effort improved the knowledge and tools for genetically manipulating *Gluconacetobacter*. Nevertheless, *E. coli* is one of the most studied microorganisms in science and is easily scalable for industrial, low-cost production. Therefore, the team decided to introduce the bioparts from *Gluconacetobacter*, needed for the bacterial cellulose production, into *E. coli*.

We will focus on one of the four genes that belong to the operon (group of genes controlled by the same promoter sequence) that is responsible for bacterial cellulose production. A promoter is a piece of DNA sequence that defines where transcription of a gene by RNA polymerase begins. The methodology explained here can be applied to any biopart you are interested in. We chose the acsD gene, which is vital for control of cellulose crystallinity and high production level (see the review in *Macromolecular Bioscience*).

The first step is finding the gene's sequence in the public database:

1. Type X54676 (accession number of the complete operon used by the iGEM team) in GENBANK's search bar. You will find several results belonging to protein sequences, scientific literature about this operon, etc. You need to choose the "Nucleotide" option.
2. Within the GENBANK entry, look for the acsD gene (see Figure 2-2).

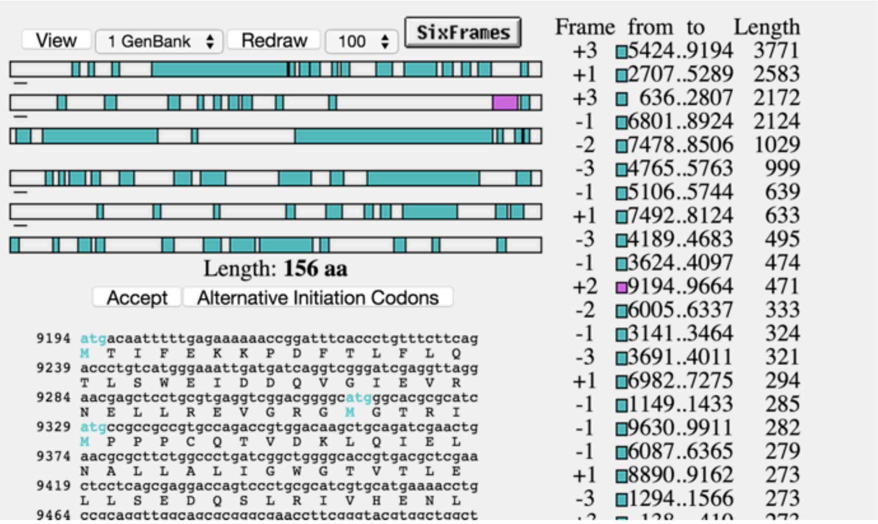


```
gene            9194..9664
                /gene="acsD"
CDS             9194..9664
                /gene="acsD"
                /codon_start=1
                /transl_table=11
                /protein_id="CAA38490.1"
                /db_xref="GI:455537"
                /db_xref="GOA:P37719"
                /db_xref="InterPro:IPR022798"
                /db_xref="PDB:3A8E"
                /db_xref="PDB:3AJ1"
                /db_xref="PDB:3AJ2"
                /db_xref="UniProtKB/Swiss-Prot:P37719"
                /translation="MTIFEKKPDFTLFLQTLSWEIDDQVGIEVRNELLREVGRGMGTR
                IMPPPCQTVDKLQIELNALLALIGWGTVTLELLSEDQSLRIVHENLPQVGSAGEPSGT
                WLAPVLEGLYGRWVTSQAGAFGDYVVTRDVDAEDLNAVPRQTIIMYMRVRSSAT"
```

*Figure 2-2. GENBANK format showing the acsD gene*

Figure 2-2 shows part of the amino acid sequence codified by the acsD gene and the nucleotide coordinates of this gene in the operon DNA (9194-9664). The easiest way to extract the acsD sequence from the complete operon sequence is to use the ORF-Finder tool. It will also help you to be sure about recovering the complete ORF (open reading frame) from the start to stop codon. Complete the following task:

1. Go to the top of the X54676 GENBANK page and select the FASTA display format.
2. Paste it on the ORF-Finder tool. Hit the OrfFind button.
3. The analysis will show you six possible frames (three forward frames and three reverse frames). Select the ORF that is in the acsD coordinates (9194-9664) by clicking the appropriate square for the frame of interest on the right-hand side: “+2 □ 9194..9664 471” (see Figure 2-3). It will also change color also on the diagram of the left-hand side.
4. Accept the ORF. Once again it will change color, indicating that it is the accepted one.



```
9194 atgacaattttgagaaaaaaccggatttcaccctgtttcttcag
     M  T  I  F  E  K  K  P  D  F  T  L  F  L  Q
9239 accctgtcatgggaaattgatgatcaggtcgggatcgaggttagg
     T  L  S  W  E  I  D  D  Q  V  G  I  E  V  R
9284 aacgagctcctgcgtgaggtcggacggggcatgggcacgcgcatc
     N  E  L  L  R  E  V  G  R  G  M  G  T  R  I
9329 atgccgccgccgtgccagaccgtggacaagctgcagatcgaactg
     M  P  P  P  C  Q  T  V  D  K  L  Q  I  E  L
9374 aacgcgcttctggccctgatcggctggggcaccgtgacgctcgaa
     N  A  L  L  A  L  I  G  W  G  T  V  T  L  E
9419 ctcctcagcgaggaccagtccctgcgcatcgtgcatgaaaacctg
     L  L  S  E  D  Q  S  L  R  I  V  H  E  N  L
9464 ccgcaggttggcagcgcgggcgaaccttcgggtacgtggctggct
```

| Frame | from to | Length |
|---|---|---|
| +3 | 🟦5424..9194 | 3771 |
| +1 | 🟦2707..5289 | 2583 |
| +3 | 🟦 636..2807 | 2172 |
| -1 | 🟦6801..8924 | 2124 |
| -2 | 🟦7478..8506 | 1029 |
| -3 | 🟦4765..5763 | 999 |
| -1 | 🟦5106..5744 | 639 |
| +1 | 🟦7492..8124 | 633 |
| -3 | 🟦4189..4683 | 495 |
| -1 | 🟦3624..4097 | 474 |
| +2 | 🟪9194..9664 | 471 |
| -2 | 🟦6005..6337 | 333 |
| -1 | 🟦3141..3464 | 324 |
| -3 | 🟦3691..4011 | 321 |
| +1 | 🟦6982..7275 | 294 |
| -1 | 🟦1149..1433 | 285 |
| -1 | 🟦9630..9911 | 282 |
| -1 | 🟦6087..6365 | 279 |
| +1 | 🟦8890..9162 | 273 |
| -3 | 🟦1294..1566 | 273 |

*Figure 2-3. ORF-Finder format showing the acsD gene open reading frame*

5. Select "2 Fasta nucleotide" instead of "1 GenBank" and press the View button. You will get the exact sequence of the acsD gene in FASTA format.

NOTE

This exercise could also have been done with the acsC, another of the operon genes. Incidentally, it would have been a bad choice because there is a mistake in the GENBANK sequence at 5286 position (it has a G instead of the correct A). Finding errors in the databases is not uncommon; this is why we recommend the use of the ORF-Finder tool, which will help you to be sure about your sequence and helps to detect this kind of problem in the online sequences.

NOTE

Do not forget that the ORF-Finder does not predict intron sequences. An intron is the nucleotide sequence within a gene that will not code for protein. In eukaryotic organisms, it is removed before translation of messenger RNA by the splicing process. If you have intron sequences, you can use a related gene/protein sequence and one of the tools.

Now that you have obtained the exact sequence of your biopart, you will need to have a company synthesize your DNA, or you can amplify out of the genome with primers and clone into a bacterial expression plasmid. For synthesized, you can ask the company to send you the DNA in a plasmid for expression, which means that the plasmid must have the other genetic elements needed for expressing the gene, such as a promoter or RBS (the ribosome binding site is the RNA sequence that must be found in mRNA for ribosomes to bind precisely and initiate translation).

Furthermore, you can modify the DNA sequence to fine-tune it for your purpose. Imperial's team optimized codons for expression in *E. coli* and used a strong RBS sequence found in the registry of parts database (BBa_B0034). They also tuned this RBS part using the Salis-Lab RBS calculator.

There are several tools to help you to modify and explore your sequence (examine the restriction enzymes profiles, insert mutations, or change the codon usage). These are graphical cloning and design tools:

- Gene Designer
- ApE
- Genome Compiler
- SnapGene
- Benchling

When you receive your biopart, you will probably want to continue with further cloning. We recommend you to plan in advance which strategy you would like to use (Gibson assembly, restrictions enzymes and ligase, etc.). We also recommend that you find which restriction enzymes don't cut your sequence. This will allow you to add these sites using primers in a PCR reaction before cloning or to ask to the synthesis company to put these sequences in your DNA. You can use NEBcutter2 and Webcutter 2.0 to analyze the restriction sites.

Both have easy-to-follow instructions. It is likely that one of them is more advantageous than the other for your specific problem. You may also be happy with using the restriction site analysis that is included in one or more of the cloning tool software packages that are listed above.

## Enhancing Your Capacity for Engineering Life

As the work of a synthetic biologist develops, it becomes less common to try to modify one gene at a time. As the field is growing fast, the community naturally aims at tuning dozens of genes at a time to get a desired function, while some would even want to attempt a complete genome change at once. To carry out those aims, you will need to be able to automate complex tasks into scripts that can be invoked by simple commands with little effort using command-line interfaces (CLIs). This allows for efficient analysis of large datasets while ensuring a low error rate. By "large datasets," we mean, for example, the ones resulting from the 1000 Genomes Project, which has so far produced 200 terabytes; or the ones produced by the Cancer Genome Project. For many bioinformatics applications, there is no choice but to use the command line, and it is unlikely that their authors will ever develop alternatives. Therefore, it is a good idea to be ready for that and acquire some skills to use CLI.

The use of CLI allows you to automate many tasks at your local or remote computer, such as downloading genes from the GENBANK, which already has this option implemented in the Entrez Direct tool application program interface (API). Once the tool is in a machine that you can interact with via a shell, you can download a copy of the bacterial cellulose production operon for local usage by typing a single command:

```
esearch -db nucleotide -query "X54676" | efetch -format fasta > mySeq
```

This operation will fetch a FASTA file with the nucleotide sequence in one go. `esearch` will get the GENBANK entry, a pipe (|) will cause `efetch` to pick the output and format it as FASTA, and the result will be "thrown" by redirection (>) into a file named *mySeq* in your current directory. Many variants are, of course, possible. If you want the GENBANK file format, use (`-format gb`) instead of (`-format fasta`). The GENBANK file format contains further information about the selected operon; for example, it contains the coordinates of each gene that composed it.

This type of operation can be scaled up to retrieve complete genomes, for example. Chaining such operations into complex scripts that perform a variety of analyses, prepare formatted outputs, submit jobs to remote servers, and a huge variety of operations will allow you to increase your capacity for engineering life beyond the limits imposed by simple interactive bioinformatic tools.

## Conclusion

The wealth of data available online today is an invaluable resource for biohackers to tinker with. Therefore, having a minimal working knowledge base of bioinformatics is a powerful asset for designing and hacking biological systems. Whether it is RNA or protein folding, sequence analysis, primer design, or restriction enzyme analysis, there's bound to be a bioinformatics tool for the job, as we've seen with our walkthrough of Imperial College's bacterial cellulose project. We hope this introduction helped you through your first steps in bioinformatics or resolved any doubts you might have had. We also hope to have triggered your curiosity and stimulated you to acquire more elaborate bioinformatics skills to use in ambitious synthetic biology projects. Now, go hack some DNA!

## Join Our Leukippos Community

If you like synthetic biology and bioinformatics, want to learn more, get some help, and be part of some great biohacking projects, join Leukippos, a synthetic biology lab in the cloud. We have an awesome group on Facebook. You are very welcome to join.

## Donations

Our Leukippos community needs your financial support. We are all working for free and pay everything out of our own pockets, such as server costs and page registrations. If you like our work and wish that we produce more quality content, such as this paper, you can donate some bitcoins to this address: 1KnikzSG7fnRfG76DxjLZyrbvw8fS9nisw.

---

*Correspondence can be directed to Dr. Gerd Moe-Behrens: leukipposinstitute@googlemail.com.*

*Edgar Andrés Ochoa Cruz (aka Don) has a PhD in biotechnology from São Paulo University, Brazil. He was the instructor of two Brazilian iGEM teams and founded Syntechbio, the first biohacker space in South America. He is managing and developing several projects in synthetic biology and developing platforms for providing home and industry users access to biotechnology. He is the cofounder of Arcturus BioCloud in San Francisco, California, taking synthetic biology to your home safely.*

*Sayane Shome completed her undergraduate studies in bioinformatics from Vellore Institute of Technology in Vellore, India. She works as an external student researcher and virtual classroom trainer for bioinformatics modules at King Abdulaziz University in Rabigh, Saudi Arabia. She served as the president of RSG-India, a student body affiliated to the International Society of Computational Biology (ISCB) Student Council, from 2012 to 2014.*

*Pablo Cárdenas is an undergrad student at Universidad de Los Andes in Bogotá, Colombia. He has been a volunteer at the Leukippos Institute since 2012 and is interested in biology, biotech, and the ethics involved in these areas (as well as all of science) in the context of improving public health and social awareness. He is an enthusiastic supporter of open science and glad to participate in projects like Leukippos.*

*Maaruthy Yelleswarapu is a master's student at ETH Zurich and has been collaborating at the Leukippos Institute since 2013.*

*Jitendra Kumar Gupta works as a programmer and research assistant at Shodhaka Life Sciences Private Limited, incubated within the Institute of Bioinformatics and Applied Biotechnology. He has a master's degree in bioinformatics from Mangalore University. For a short period of time, he worked in the Manipal Institute of Technology as a research scholar. He joined the Open Source Drug Discovery Project (an initiative by CSIR) at the Indian Institute of Science.*

*Eugenio Maria Battaglia is an undergraduate student in molecular biotechnology at the University of Turin with a specialization in integrative neuroscience. Currently he's developing the concept of a bio-commons license in the European framework named Synenergene.*

*Alioune Ngom received his BSc degree in mathematics and computer science from the Universite du Quebec a Trois-Rivieres in 1990 and his MSc and PhD degrees in computer science from the University of Ottawa in 1994. He is a professor at the University of Windsor, Ontario, Canada. Prior to joining the University of Windsor, he was an assistant professor at the Department of Mathematics and Computer Science at Lakehead University, Thunder Bay, Ontario, Canada, from 1998 to 2000. During his short stay at Lakehead University in 1999, he cofounded Genesis Genomics Inc. (now, Mitomics Inc.), a biotechnology company specializing in the analysis of the mitochondrial genome and the identification and design of mtDNA biomarkers for the early detection of cancer. He is member of the IEEE-BBTC and IAPR-Bioinf and coleads the Pattern Recognition in Bioinformatics group at the University of Windsor.*

*Pedro L. Fernandes is a bioinformatics training coordinator at Instituto Gulbenkian de Ciência, in Oeiras, Portugal. He also organizes brainstorming events on challenging themes such as "Distance and eLlearning Technologies," "Systems Biology and P4 Medicine," and "Pathway Analysis in Proteomics." He is an advisor to Figshare, an ambassador to iAnn, and plays team-leading roles in organizations like EMBnet and GOBLET.*

*Gerd Moe-Behrens has a BSc and MSc from the University of Oslo and a PhD from the Faculty of Medicine, University of Oslo, Norway. He founded the Leukippos Institute for Synthetic Biology, a research institute solely in the cloud. Moreover, he is founder and CEO of CytoComp, a young startup focusing on biological computing.*