DM536 Introduction to Programming
Fall 2014 Project
/
DM550 Introduction to Programming
Fall 2014 Project (Part 1)

Department of Mathematics and Computer Science
University of Southern Denmark

October 24, 2014

Introduction
The purpose of the project for DM536 and DM550 Part 1 is to try in practice the use of programming techniques and knowledge about the programming language Python on small but interesting examples. There are two possible projects. You have to pick one of these.

Please make sure to read this entire note before starting your work on this project. Pay close attention to the sections on deadlines, deliverables, and exam rules.

Exam Rules
This project is the final exam for DM536 and it is part of the final exam for DM550. For DM550 two projects have to be passed to pass the overall exams. Thus, the project must be done individually, and no cooperation is allowed beyond what is explicitly stated in this document.

Deliverables
A short project report (at least 6 pages without appendix) has to be delivered. This report has to contain the following 6 sections:

- **front page** (course number, name, section, date of birth)
- **specification** (what the program is supposed to do)
- **design** (how the program was planned)
- **implementation** (how the program was written)
- **testing** (what tests you performed)
- **conclusion** (how satisfying the result is)

The report has to be delivered as a PDF file (suffix .pdf) together with the Python source code files (suffix .py) electronically using Blackboard's SDU Assignment functionality. Do not forget to include the complete source code!

Deadlines

The description of the projects indicates which tasks have to be solved by the pre-delivery deadline (1 & 2 for Fractals project; 1, 2 & 3 for DNA project). All tasks except for the optional tasks (3 & 9 for Fractals project; 4 & 9 for DNA project) have to be solved by the final delivery deadline.

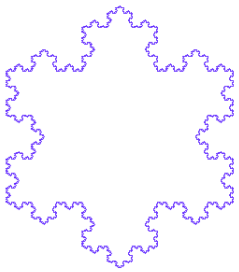
Pre-Delivery:	Thursday, October 2, 23:59
Final Delivery:	Thursday, October 30, 23:59

Project "Fractals and the Beauty of Nature"

Fractals are geometric objects that are similar to themselves on arbitrarily small scales. There are many examples of fractals in nature, and they form some of the most beautiful structures you can find. One example is snowflakes, but also lightning has a fractal structure. Another example are ferns, where each part is similar to the whole.

In this project, we will use Swampy to generate fractals that are similar to structures found in nature.

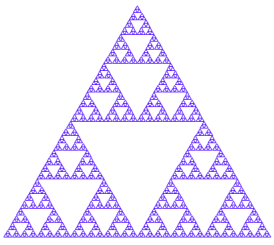
Task 0 – Preparation I



On the course home page you find an example using a Koch curve to render a snowflake (`koch.py`, `FractalWorld.py`, `koch2.png`, and `koch5.png`). Use this example and experiment with the depth and the length parameters. Make sure you use the updated file `FractalWorld.py`. Understand what (new) features are used here. In particular, line width and color options have been added to `fd`.

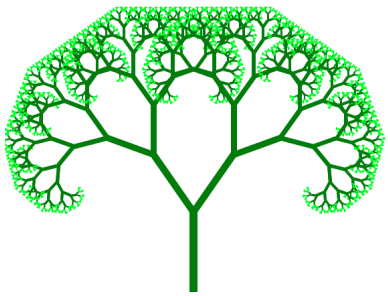
Task 1 – Sierpinski Triangle (include in Pre-Delivery)

A Sierpinski triangle is a triangle divided into three triangles, that are demselves divided into three triangles etc. On the course home page, you find four examples (`sierpinski0.png`, `sierpinski1.png`, `sierpinski2.png`, and `sierpinski5.png`). On the right-hand side you see a Sierpinski triangle of depth 5, i.e., a triangle that has been subdivided five times.



You task is to write a Python program that draws a Sierpinski triangle. To this end, you can follow the approach of the Koch Snowflake, i.e., at depth 0 you draw a triangle. At any other depth n, you draw Sierpinski triangle of depth $n-1$, move to the next corner, draw another Sierpinski triangle of depth $n-1$, move to the last corner, draw the third Sierpinski triangle of depth $n-1$, and, finally, move to the original corner.

Task 2 – Binary Tree (include in Pre-Delivery)

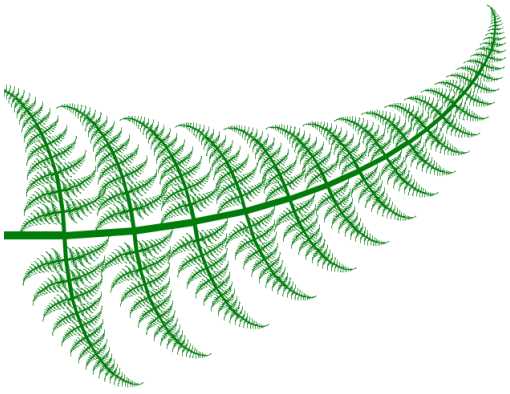


A binary tree is a tree, which is obtained by first drawing a stem and then subdividing into two branches. A tree consisting of just a stem is called a tree of depth 1. The picture to the left shows a tree of depth 12. On the course home page, you find examples (`tree3.png`, `tree6.png`, `tree9.png`, and `tree12.png`). Here, as an addition, the last two layers of branches have been colored `darkgreen` instead of `green`.

You task is to write a Python program that draws a binary tree. To this end, you have to modify the approach from the previous tasks slightly. The main idea is to draw a line, then turn and draw the left branch, then turn and draw the right branch, and, finally, go back where you came from.

Task 3* – Fern Time

There are three differences between a fern and a binary tree. First, the fern has three branches from each stem and each subbranch. Second, the fern use some small constant angle to turn the middle branch. Third, the middle branch is scaled down less than the left and the right branch. On the course home page, you find examples (`fern03.png`, `fern06.png`, `fern12.png`, and `fern24.png`).



Your challenge task is to write a Python program that draws a fern. To this end, you have to modify the approach from the previous task. The main idea is that due to the third difference, a constant recursion depth is not useful in implementing a fern. Instead, define a limit for the size of the subbranches drawn and use this as the base case of the recursion.

Note that this task is optional and does not have to be solved for this project to be considered as passed.

Fractal Description Language

Many fractals can be generated from descriptions in the Fractal Description Language (.fdl). This language describes a start state of a fractal and a set of rules how to progress to larger depths.

For example, to generate a Koch curve, we start with the initial (depth 0) state `F` signifying a forward move, i.e., a straight line. The rule for expanding to the next depth is given as a replacement rule.

```
F -> F L F R F L F
```

where `L` is a 60 degree turn to the left and `R` is a 120 degree turn to the right. That is, we replace a straight line by a straight line, a left-turn, a straight line, a sharp right-turn, a straight line, a left-turn, and a fourth and final straight line.

Thus, the state for depth 1 is `F L F R F L F`. To get to depth 2, we have to apply the rule again to all positions of the state where it is possible, i.e., we have to replace each of the four `F` by `F L F R F L F`. The result is F L F R F L F L F L F R F L F R F L F L F L F R F L F R F L F L F L F R F L F where the new sections are underlined to aid your understanding.

To get to depth 3, we would have to replace each of the 16 `F`s by the right side of the rule. For the sake of brevity, I leave this exercise to you.

Let us take a look at the file `koch.fdl` available from the project section of the course home page.

```
start F
rule F -> F L F R F L F
length 2
depth 5
cmd F fd
cmd L lt 60
cmd R rt 120
```

The first line give the start state, i.e., the state `F` for depth 0. The second line give the only rule needed for the Koch curve, i.e., to replace `F` by `F L F R F L F`. The third line specifies the length of each segment, i.e., each straight line will be 2 units long. The fourth line specifies that states should be expanded to depth 5 before drawing the fractal. Finally, the Lines 5 to 7 specify that `F` is a straight line, `L` is a 60 degree left-turn and `R` is a 120 degree right-turn.

Task 4 – Preparation II
Download all the .fdl files from the course homepage (at least `dragon.fdl`, `fern.fdl`, `koch.fdl`, `sierpinski.fdl`, `sierpinski2.fdl`, `snowflake.fdl`, and `tree.fdl`).

Try to understand how these generate their respective fractals. While most of the fractals are known from earlier tasks, `dragon.fdl` represents a dragon curve and `sierpinski.fdl` represents a Sierpinski curve.

Task 5 – Representing and Applying Rules
A rule consists of a single letter for the left side and a list of letters for the right side. Your task is to create a user-defined type (= Python class) "Rule" that represents such a rule. This class should have attributes for at least the left side and the right side.

You also need to write a function that applies a rule to each (matching) element of a list, i.e., that from the list `["F","L","F","R","F","L","F"]` for depth 1 of the Koch curve will generate the list `[["F","L","F","R","F","L","F"],"L",["F","L","F","R","F","L","F"],"R",["F","L","F","R","F","L","F"],"L",["F","L","F","R","F","L","F"]]`.

Task 6 – Representing and Executing Commands
A command consists of a command string (such as `fd`, `lt`, `rt`, `scale`) and a list of arguments. Your task is to create a user-defined type (= Python class) "Command" that represents a command. This class should have attributes for at least the command string and the arguments.

You also need to write a function for executing a command. This function gets as an argument a turtle object and a length. The command with command string `lt` and argument list ["60"] should execute the Python statement `lt(turtle, 60)`. The command `scale` multiplies the length with the given float. Finally, the command `nop` simply does nothing (no operation).

Task 7 – Loading a Fractal Description Language File
A fractal consists of a start state represented by a list, a list of rules, a mapping from single letters to commands, a length, and a depth.

Your task is to create a user-defined type (= Python class) "Fractal" that has at least the attributes described above. In addition, you should write a function or method that executes all commands of a given state, i.e., goes through a state list, uses the mapping from single letters to commands, and executes these.

You also need to write a function that reads an .fdl file and creates a Fractal object from it.

Task 8 – Generating Fractals
Now you are left with computing a new state from an old state. Your task is to write a (recursive or iterative) function or method that for a given start state, set of rules, and depth computes the state at that depth.

Finally, you have to put everything together such that you can load, compute, and draw a fractal for a given file. To this end, you should import the `sys` module and use as a filename the first argument passed on the command line, i.e., `sys.argv[1]`. You also need to write a function to flatten the lists of lists obtained by rule applications into a list of elements those lists.

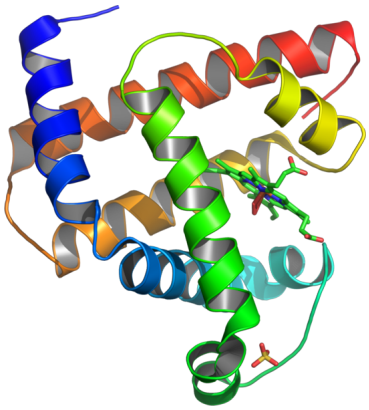
Task 9* – Support for Line Widths and Colors
The fractals look nice enough, but some colors and wider lines would make them more pretty. Your challenge task is to extend the Fractal Descripton language by the commands `color` and `width` where color gets a color name, a color code or "`random`" as an argument while width gets a float. Random colors can e.g. be generated by using format strings:
`"#%02x%02x%02x" % (randint(0,255),randint(0,255),randint(0,255))`

Here is an example for a nicer dragon curve:

```
start F X
rule X -> X R Y F
rule Y -> F X L Y
length 3
depth 13
color random
width 2.0
cmd F fd
cmd X nop
cmd Y nop
cmd L lt 90
cmd R rt 90
```

Note that this task is optional and does not have to be solved for this project to be considered as passed.

Project "From DNA to Proteins"



In nature, deoxyribonucleic acid (short: DNA) is used to encode genetic information of living organisms as sequences of bases. There are four bases found in DNA: adenine (short: `A`), cytosine (short: `C`), guanine (short: `G`), and thymine (short: `T`). One of the main functions of DNA is to encode the sequence of amino acids used in the construction of proteins.

Modern technological advances have made it possible to decipher this genetic information. You can find for example the base sequences of human chromosomes on the following web site:

`http://hgdownload.cse.ucsc.edu/downloads.html`

In this project, we will assemble base sequences from such files and analyze these sequences to identify proteins.

Task 0 – Preparation I

Download the file `hg38.chromFa.tar.gz` from the full data set for human and unpack it. Locate the file `chrX.fa` and view it in a text viewer or editor.

The first line is a short description of the sequence contained in the file. The rest of the lines is the base sequence. You will find lower-case and upper-case variants of the bases `A`, `C`, `G`, and `T`. In addition you will find `N`, which for the purpose of this project can be ignored.

Task 1 – Assembling the Sequence (include in Pre-Delivery)

For our purposes, we will ignore `N`s and we will ignore the difference between lower-case `a`, `c`, `g`, `t` and upper-case `A`, `C`, `G`, `T`.

Your task is to write a program that reads all lines and constructs one long string containing the base sequence. In this process, whitespace and `N`s have to be ignored. In addition, all base pairs should be represented by upper-case letters. Use the smaller file to print the result and compare the beginning of the assembled sequence manually to the original file.

Task 2 – Finding Starting Points (include in Pre-Delivery)
The construction of a protein by a ribosome begins at a start codon. On our DNA, this is denoted by the base sequence `ATG`. This is called a *start codon*. Approximately 25 bases earlier in the sequence we often find a so-called TATA box. This is a base sequence `TATAAA`.

Your task is to write a function that will find all positions of a start codon that occurs 15-30 bases after the beginning of a TATA box. To this end, first write a function that will find all positions of the string `TATAAA` in the sequence. Then, find out how far the next ATG is located. If the distance is inside the admissible range, remember the index of the start codon.

Task 3 – Finding End Points (include in Pre-Delivery)
Each protein is encoded by a sequence of bases starting with ATG. Every amino acid is encoded by three bases. The end of a protein is given by a *stop codon*. This can be any of the following three sequences: `TAG`, `TAA`, or `TGA`.

Your task is to write a function that will identify the end point of a protein, i.e., the index of the first stop codon encountered after the start codon. To this end, you need to advance in steps of 3 through the base sequence until you encounter either the end of the sequence or a stop codon.

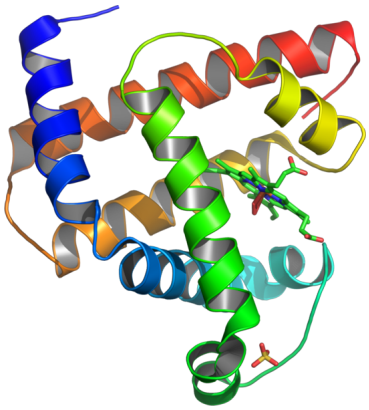
Task 4* – Potential Proteins without TATA Boxes
Not all proteins are prefixed by a TATA box occurring 15-30 bases earlier. In general, potential proteins can be overlapping. For example, the base sequence `ATGAATGAATAGATGA` contains a potential protein at index 0 (`ATGAATGAATAG`) and at index 4 (`ATGAATAGATGA`). We call this a genuine overlap. There is also trivial overlap, when `ATG` occurs inside a base sequence at a position divisible by three, e.g., `ATGATGTAG`.

Your challenge task is to identify all potential starts of proteins and answer the following two questions w.r.t. the file `chrX.fa`:

- How many genuine overlaps are there?
- How many potential proteins are there? Here, for genuine overlaps both proteins count while for trivial overlaps, only the longest protein counts.

Note that this task is optional and does not have to be solved for this project to be considered as passed.

Amino Acids



In nature, deoxyribonucleic acid (short: DNA) is used to encode genetic information of living organisms as sequences of bases. There are four bases found in DNA: adenine (short: `A`), cytosine (short: `C`), guanine (short: `G`), and thymine (short: `T`). One of the main functions of DNA is to encode the sequence of amino acids used in the construction of proteins.

In the following tasks, we will translate the base sequences identified in the previous tasks into proteins. You find examples of output on the course home page (`chr1_gl000191_random.pro` and `chrX.pro`).

Task 5 – Preparation II
Start by modifying your program from the previous tasks to produce the list of base sequences that you found. This should be easy as you already compute the start and end index of theses substrings. Do not forget to skip the start codon!

Task 6 – Representing Amino Acids
Download the file `dna-codons.cdl` from the course home page. It encodes the genetic codes for the twenty amino acids. The description for one amino acid starts with a "`>`" followed by its abbreviation, its short name, and its long name. In the following lines until the next amino acid, each line identifies a codon that is translated to this amino acid.

Your task is to create a user-defined type (= Python class) "Acid" that represents one amino acid. This class should have attributes for at least the abbreviation, the short name, the long name, and the codons that encode this amino acid. For example, for aspartic acid this would be the abbreviation "D", the short name "Asp", the long name "Aspartic acid", and a list of codons containing exactly "GAT" and "GAC". Write a function that will read `dan-codons.cdl` and produce a list of 20 amino acids.

Task 7 – Setting up the Translation
As we will translate a large number of base sequence to proteins, it is important to have an efficient translation from codons to amino acids.

Your task is to create a user-defined type (= Python class) "Ribosome" that builds and stores a mapping of codons to amino acids. To this end, your

type should have at least two associated functions or methods. First, you need a function that takes an amino acid (represented as in Task 6) and adds mappings from its codons to the amino acid. Second, you need a function that takes a codon (a base sequence of length 3) and returns the appropriate amino acid, i.e., the amino acid that this codon is translated to.

Task 8 – Creating Proteins
A protein is represented by a sequence of amino acids.

Your task is to create a user-defined type (= Python class) "Protein" that builds and stores sequences of amino acids. To this end, your type should have at least two associated functions or methods. First, you need a function that takes a base sequence and uses the function associated with the "Ribosome" class to translate it into a sequence of amino acids represented by instances of the "Acid" class. Second, you need a function that will convert a protein into a string. This function should have a parameter "mode" where a value of "0" means that a string of one-letter abbreviations is returned (e.g. `ADDYF`), a value of "1" means that a string of comma-separated short names is returned (e.g. `Ala, Asp, Asp, Tyr, Phe`), and a value of "2" means that a string of new-line separated long names is returned, e.g.:

```
Alanine
Aspartic acid
Aspartic acid
Tyrosine
Phenylalanine
```

Task 9* – Representing Codons
Codons are base sequences of length 3 that are encoded into exactly one amino acid. Your challenge task is to create a user-defined type (= Python class) "Codon" that represents a base sequence of length 3 and to use it instead of using strings in Tasks 5–8. You have to write a function or method, that takes a base sequences from Task 5 and translates it into a list of instances of "Codon". You will also need to write your own `__hash__(self)` and `__cmp__` methods in order to be able to use codons as keys for a dictionary.

Note that this task is optional and does not have to be solved for this project to be considered as passed.