What the APL is a k-mer?

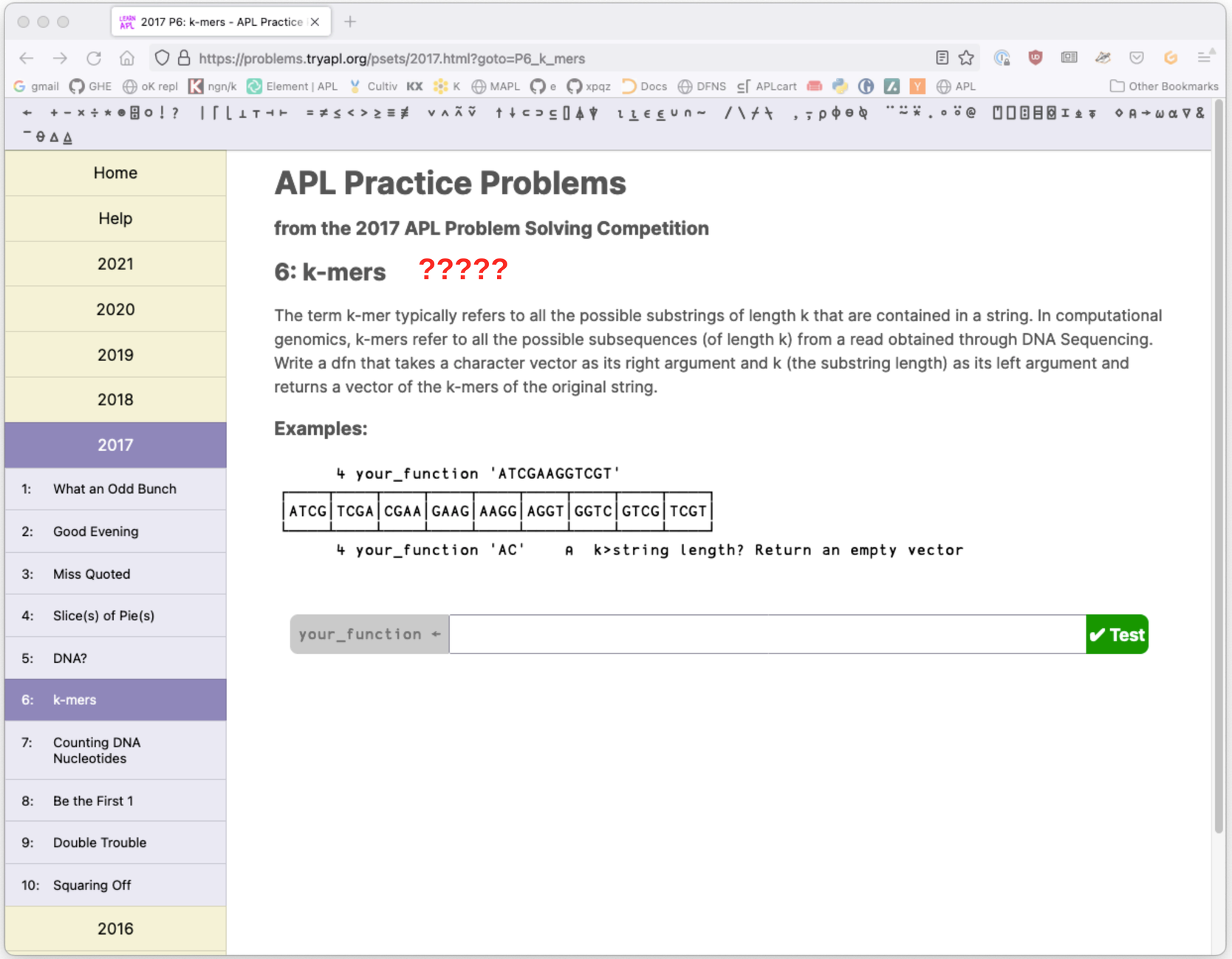
Stefan Kruger

I write stuff on array languages

xpqz.github.io/learnapl

xpqz.github.io/cultivations

xpqz.github.io/kbook


2017 P6: k-mers - APL Practice
https://problems.tryapl.org/psets/2017.html?goto=P6_k_mers
Home
Help
2021
2020
2019
2018
2017
1: What an Odd Bunch
2: Good Evening
3: Miss Quoted
4: Slice(s) of Pie(s)
5: DNA?
6: k-mers
7: Counting DNA Nucleotides
8: Be the First 1
9: Double Trouble
10: Squaring Off
2016
APL Practice Problems
from the 2017 APL Problem Solving Competition
6: k-mers
?????
The term k-mer typically refers to all the possible substrings of length k that are contained in a string. In computational genomics, k-mers refer to all the possible subsequences (of length k) from a read obtained through DNA Sequencing. Write a dfn that takes a character vector as its right argument and k (the substring length) as its left argument and returns a vector of the k-mers of the original string.
Examples:
4 your_function 'ATCGAAGGTCGT'
ATCG TCGA CGAA GAAG AAGG AGGT GGTC GTCG TCGT
4 your_function 'AC' ⍬ k>string length? Return an empty vector
your_function ←
Test

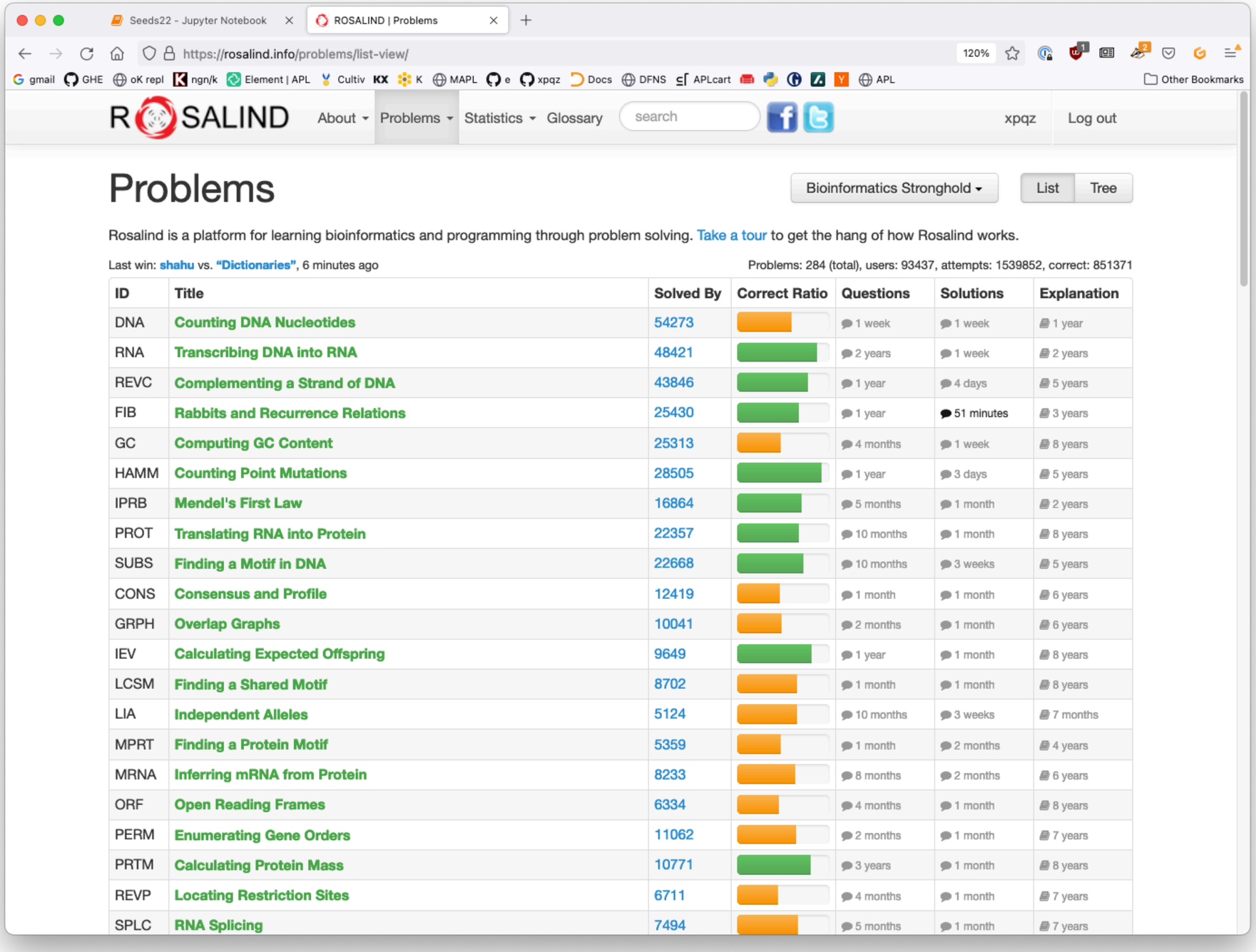
Bioinformatics?

> Computational and statistical analysis to decipher biology from genome sequences and related data, including both DNA and RNA sequence as well as other "post-genomic" data

Wikipedia

Bioinformatics?

- A DNA-string is just a vector...
- A field ideally suited to learning APL!
- ...even if you don't know your k-mers from your spliced motifs



ID	Title	Solved By	Correct Ratio	Questions	Solutions	Explanation
DNA	Counting DNA Nucleotides	54273		1 week	1 week	1 year
RNA	Transcribing DNA into RNA	48421		2 years	1 week	2 years
REVC	Complementing a Strand of DNA	43846		1 year	4 days	5 years
FIB	Rabbits and Recurrence Relations	25430		1 year	51 minutes	3 years
GC	Computing GC Content	25313		4 months	1 week	8 years
HAMM	Counting Point Mutations	28505		1 year	3 days	5 years
IPRB	Mendel's First Law	16864		5 months	1 month	2 years
PROT	Translating RNA into Protein	22357		10 months	1 month	8 years
SUBS	Finding a Motif in DNA	22668		10 months	3 weeks	5 years
CONS	Consensus and Profile	12419		1 month	1 month	6 years
GRPH	Overlap Graphs	10041		2 months	1 month	6 years
IEV	Calculating Expected Offspring	9649		1 year	1 month	8 years
LCSM	Finding a Shared Motif	8702		1 month	1 month	8 years
LIA	Independent Alleles	5124		10 months	3 weeks	7 months
MPRT	Finding a Protein Motif	5359		1 month	2 months	4 years
MRNA	Inferring mRNA from Protein	8233		8 months	2 months	6 years
ORF	Open Reading Frames	6334		4 months	1 month	8 years
PERM	Enumerating Gene Orders	11062		2 months	1 month	7 years
PRTM	Calculating Protein Mass	10771		3 years	1 month	8 years
REVP	Locating Restriction Sites	6711		4 months	1 month	7 years
SPLC	RNA Splicing	7494		5 months	1 month	7 years

Project Rosalind

- Bioinformatics problem collection: rosalind.info
- Automated results check; language independent
- No walk-over, but solution ratios indicate difficulty
- Also: strict 5 min time limit, including data download and result upload

Writing good APL

- Clarity > efficiency, where opposed
- Actual performance can depend on Dyalog version and actual processor features

APL Practice Problems

from the 2017 APL Problem Solving Competition

6: k-mers

The term k-mer typically refers to all the possible substrings of length k that are contained in a string. In computational genomics, k-mers refer to all the possible subsequences (of length k) from a read obtained through DNA Sequencing. Write a dfn that takes a character vector as its right argument and k (the substring length) as its left argument and returns a vector of the k-mers of the original string.

Examples:

```
      4 your_function 'ATCGAAGGTCGT'
┌────┬────┬────┬────┬────┬────┬────┬────┬────┐
│ATCG│TCGA│CGAA│GAAG│AAGG│AGGT│GGTC│GTCG│TCGT│
└────┴────┴────┴────┴────┴────┴────┴────┴────┘
      4 your_function 'AC'    ⍬  k>string length? Return an empty vector
```

Skipping this!!

https://problems.tryapl.org/psets/2017.html?goto=P6_k_mers