Robert McCaull

Bio 131 Final Project Report (Spring 2020)

My project was to o!i"y one o" t#e algorit# s we learne! in class (a gree!y oti"$"in!ing algorit#) so t#at it woul! sprea! its wor% across se&eral e'ecution t#rea!s(rat#er t#an just a single one) *n t#eory(by !oing t#is(a co puter coul! e'ecute certain intensi&e parts o" t#e algorit# in parallel to eac# ot#er(allowing a co puter wit# ultiple processors to "ully ta%e a!&antage o" its resources w#en e'ecuting t#e algorit#) +#is coul! potentially result in uc# "aster co putations(i" any processors were a&ailable) *n practice(t#ere are co plicating "actors to ac#ie&ing t#is(so e o" w#ic# * wasn,t able to sol&e) -owe&er(* was able to pro!uce a correct algorit# t#at (t#eoretically) ac#ie&es t#e goal o" paralleli.ation)

+#e biological proble t#at y project is !irecte! towar!s is t#e proble o" "in!ing co onalities between !i""erent / 0 1 se2uences) *" we #a&e a set o" se2uences w#ic# all ser&e a si ilar "unction(an! we #a&e t#e ability to "in! co on "actors between t#e (we can narrow in on w#ic# parts o" t#e se2uences contribute towar!s t#e co on "unction)

+o sol&e t#is biological proble (we "irst sol&e a relate! co putational proble (an! t#en apply our solution to t#e particular case we,re intereste! in) +#e relate! co putational proble is as "ollows3 gi&en a list o" strings an! an integer %(we woul! li%e to co e up wit# a list %$lengt# substrings (calle! oti"s or % ers)(one "or eac# string(w#ic# are as si ilar to eac# ot#er as possible) By 4as si ilar as possible,(we ean t#e ones w#ic# !i""er as little as possible "ro t#e consensus string t#ey enco!e)

Searc#ing t#roug# all possible lists o" oti"s to "in! t#e best one woul! ta%e a pro#ibiti&ely long ti e) +o !eal wit# t#is(we use a gree!y algorit# w#ic# only consi!ers a subset o" t#e possible lists o" oti"s) +#e lists o" oti"s t#is algorit# outputs are not guarantee! to be per"ect 5 suc# is t#e nature o" a gree!y algorit# 5 but t#ey !on,t nee! to be per"ect to be use"ul(an! t#e gains we a%e in

re!uce! running ti e are consi!erable) My project is just an e'tension o" t#is gree!y algorit# (w#ic# is (t#eoretically) "aster still(since it can be run concurrently on ultiple processors)

+#e non$parallel gree!y algorit# wor%s by starting wit# a oti" "ro t#e "irst input string(an! t#en gree!ily a!!ing t#e best oti" (gi&en t#e oti"s t#at were alrea!y c#osen) "ro eac# "ollowing string in turn(to buil! a "ull list o" can!i!ate oti"s) *t t#en repeats t#is process "or eac# ot#er oti" in t#e "irst string) 1s it !oes t#is(it %eeps trac% o" w#ic# "ull list o" oti"s it #as seen so "ar t#at was best(an! at t#e en! it returns t#a21p 3)