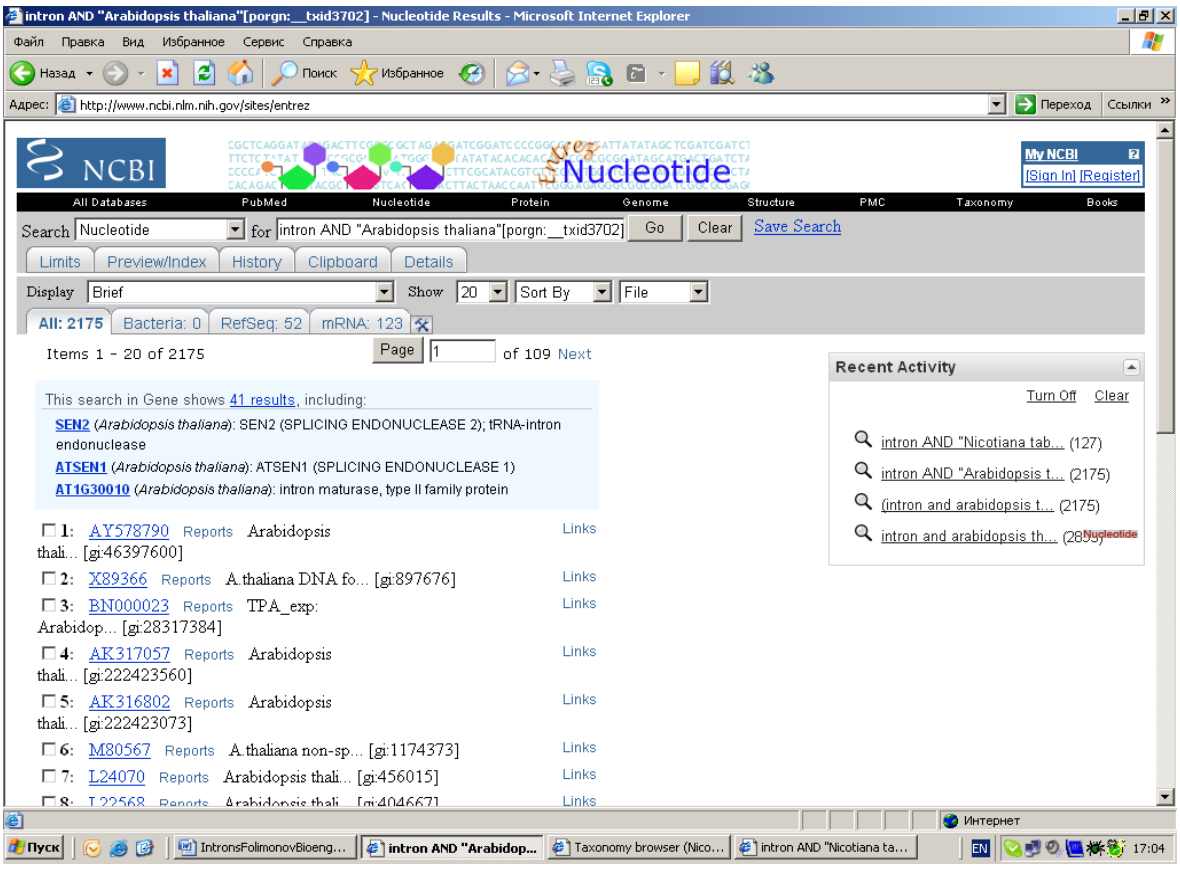
Задача для курсовой работы - написать и отладить программу из двух частей: А) создающую и обновляющую простые базы данных по нуклеотидным последовательностям интронов из заданных видов растений; Б) исследующую заданную нуклеотидную последовательность на предмет возможности вставить в нее интрон(ы) с имеющимся нуклеотидным контекстом из созданных баз.
Необходимые знания: программирование (желательно, C++), биоинформатика.

К кому обратиться:
1. Алексей Фолимонов
Старший научный сотрудник
Группа молекулярной биологии вирусов растений
Центр "Биоинженерия" РАН
Проспект 60-летия Октября, д.7, к.1
Москва, 117312
телефон +(8) 499-135-12-40
мобильный +(8) 909-672-73-64
e-mail folimonov@biengi.ac.ru

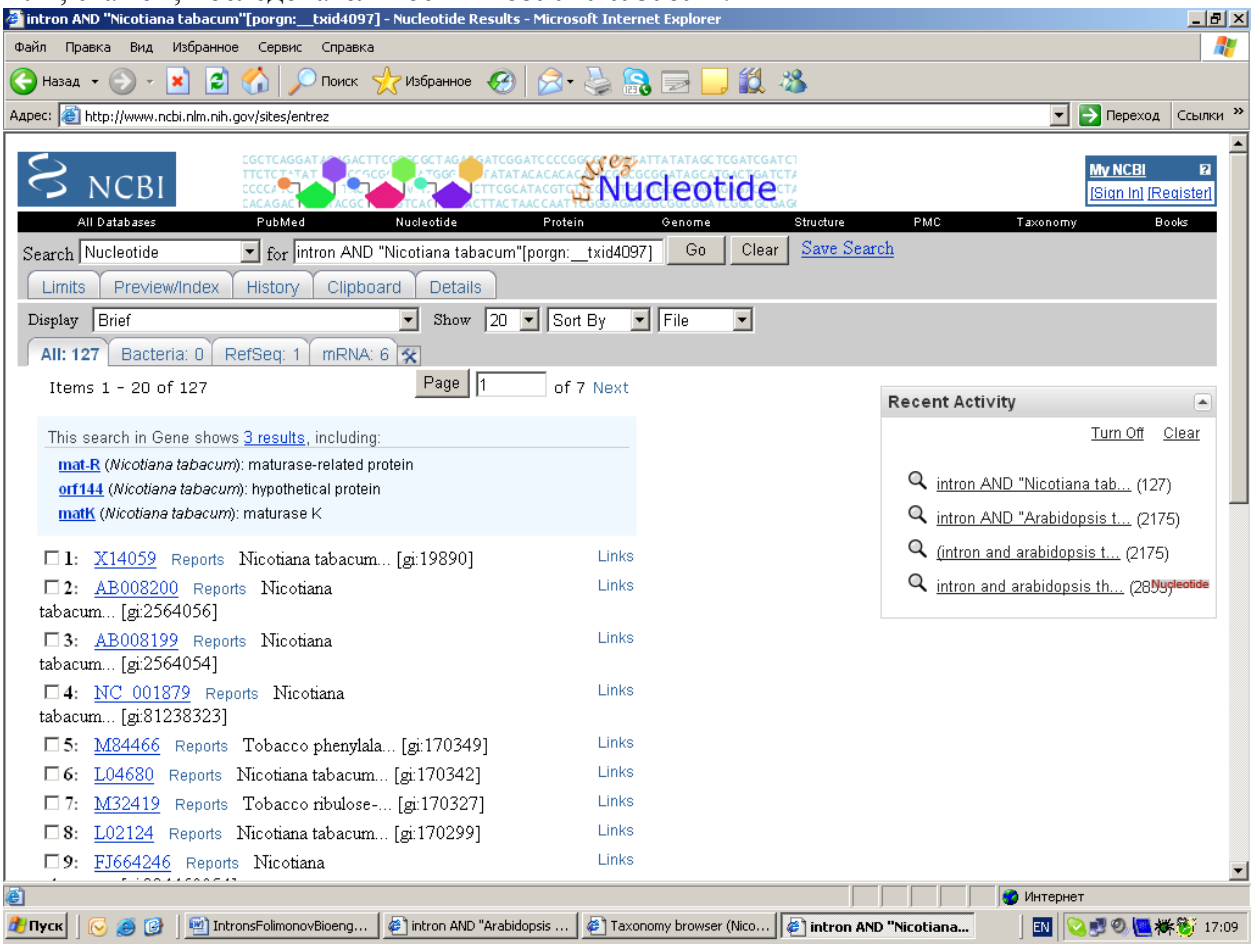
2. Алексей Анатольевич Аграновский
Профессор кафедры вирусологии Биологического факультета МГУ
телефон +(8) 495-939-23-63
e-mail aaa@genebee.msu.su

Вот, что делаю я (а должна делать программа А):

1. Найдем в ГенБанке все последовательности *Arabidopsis thaliana*, содержащие интроны:

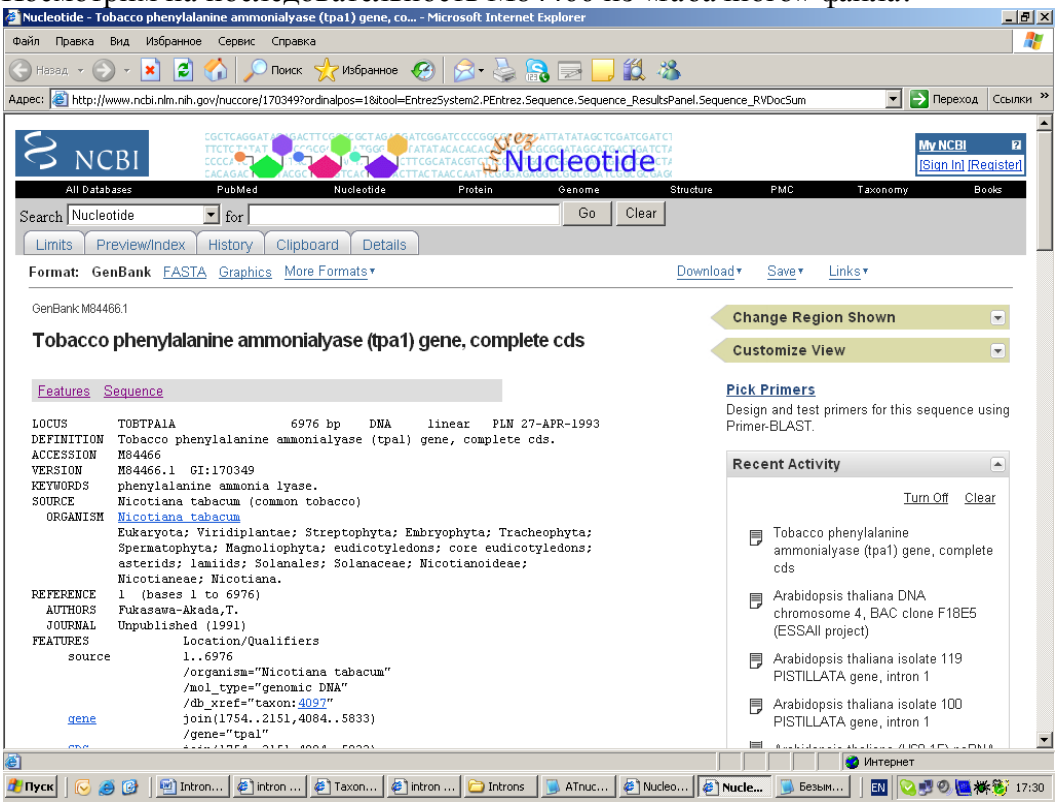


Или, скажем, последовательности *Nicotiana tabacum*:

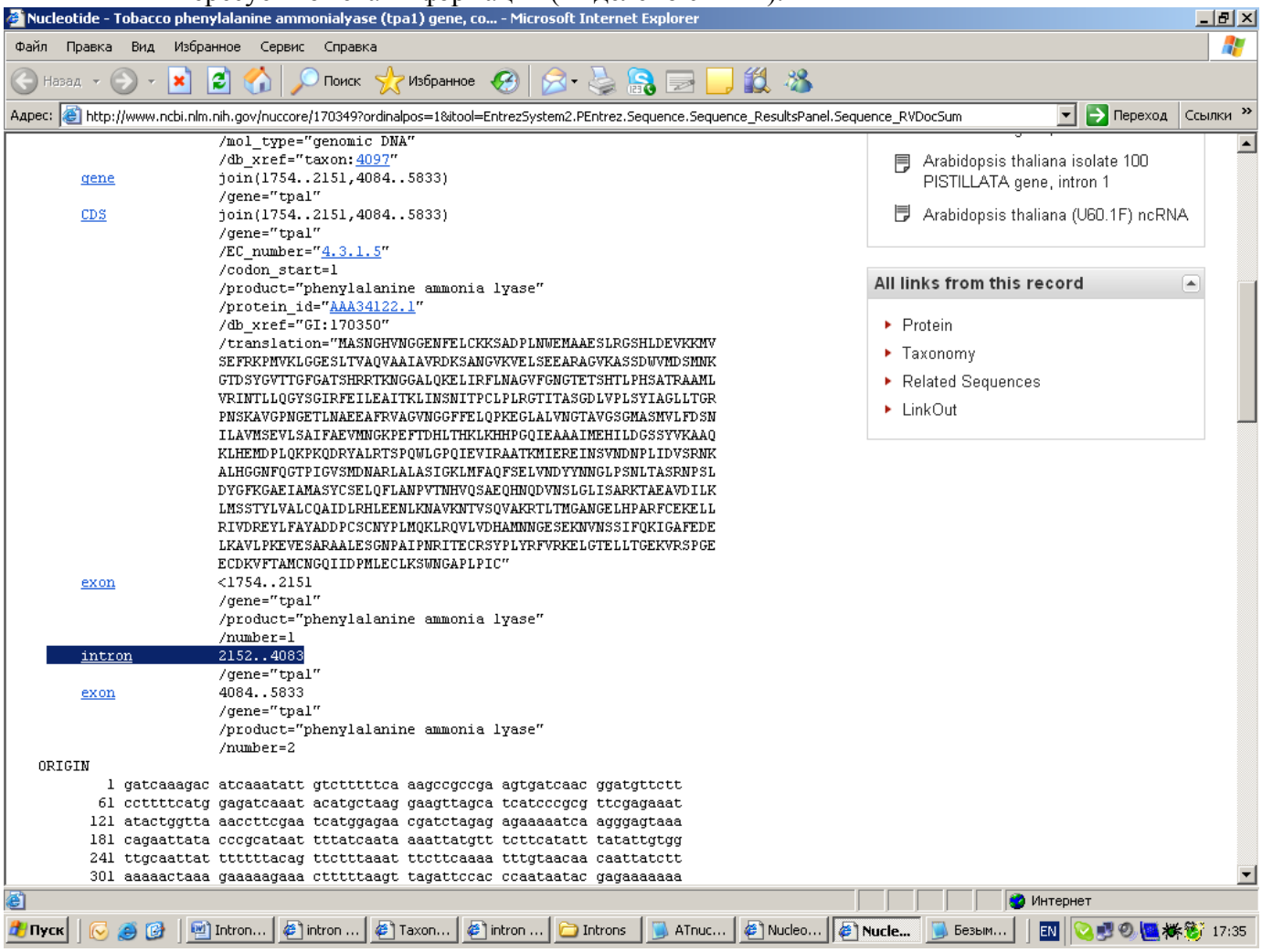


Или любого другого растения, для которого есть название и txid, хотя арабидопсис, конечно, интереснее, просто потому, что последовательностей больше…

2. Сохраним номера последовательностей в соответствующих файлах.
3. Посмотрим на последовательность M84466 из «табачного» файла:



Интересует вот эта информация (выделено синим):



4. Выделяем последовательности интрона (показано красным) и по 20 нуклеотидов 5’- и 3’-концевого экзона (показано зеленым), вот они:

cttcaaaaagaacttattag gtaaaccac
2161 ttttcatcta cttaactttg ctttctgcac cagaatttta atattttgtt tcctcctaaa
2221 gatttttatc gttaagtgtg ttaaataatt taaattaatt attttcaaga tatacgagtt
2281 tattatacac aagatttctg tcagaagttt gcagtgaccc ttagtccgaa aaaaataata
2341 gcatctacat ttttctatga cagcgttaag gaaattttac acggtcagta gaaaaggtta
2401 ttttacgtta ttatttttagt ttattgtaag cagctatata tatatatata tatatatata
2461 gttatgtttt agatggtcgt aaatttattt tgtgttgacc atgtatagta atagttaaat
2521 tcttattctg caagaaaatc tgactagcat ttatgagtta atgatctact attttgagtt
2581 aataactatt aatgatatat actaatttga ccttataaaa aaattatcta ttattttgag
2641 tcaatagtta ttaatgattt actttaactc ttatgtcaac tactcttttt gtaagaggaa
2701 atgagtttta aatgatatca actttaaggt aaaatgtgta cattttctat gtaatctaaa
2761 cgattatagc agctatattt caagttatta tttccgacca aatatgaaaa gtctcttatt
2821 attatagaac aatataatct ggtactataa atttttctta ttttatcata gtagagaaga
2881 taaagtttgt atatgtaaac atttatcaca cgcgctagac caagaactaa gtactttaga
2941 tttacatctt gaaatttgaa ttattgagta aagaaaaata ttaaatagcc cttggaaaaa
3001 tattttatat tggaacttga tacgaagtac aggaattttg catctacatc tataaatcta
3061 attatataaa ttaaattttg tacaactgta aaaaaaaaac aagccactag tttgacttt
3121 gttgggttgg aaaaagagtg tgtggttggt ggaggagatt ttagtcattt caccttcctt
3181 ctagttttgt ataaagtaca tgtaagaatt tattcagtaa atttttccta ctactatttt
3241 cctaaattta gtaagaaaaa ataatttaaa taaacatctt taaatagtta aattatttttt
3301 gtgggtcctc aaataataga tgagagacaa gtgggcttat ttaaacagtt taaatttcca
3361 actatttggg tcgaatccca gtagtcccat gtgctgtctc ttgtcagata tctatttatt
3421 ttaaagtaca aataatttttt aaaacaaata ttgcataaat aagtaatttt atacttattt
3481 aaagagaatt gttaaaaaaa ttatgaaatg acaattttaa atctcaaatt taaacttcac
3541 ttttatttttt gggcctaatt tacctttaac ctgcaccgaa actatttata tctgatagct
3601 acaaaagtgt ataaacttta tatatttttt tatacaacat acatatatat atatacaaaa
3661 atatatatat attttcgctt attatttttat tagcggctat gtcatttttt tccttttttc
3721 attttcccct cccctgccca caaggccaaa aatataacaa atgtgtgccc cactttaaag
3781 gggctttata tcacttgaaa tgatgggagt tgatgtaaca cactacaagt ttttctttt
3841 tttttcttaa gccttataat tcttgatccc tattctagtc attttctcgt actaaacaat
3901 aatatatcca atataatctc atactaattt tctcatagac cctcggctaa ttttctcata
3961 ctagtaattt ttaaatatcg tcacaaagta gataggacaa agtacaaata aaacatgttc
4021 ttatttaaat ttggttggta caattcagtt caatagctga tgttgttgac attgaattgg
4081 cag gttcttgaatgctggtgttt

5. В базе данных нам нужна информация: 1) о последовательности экзонов; 2) о последовательности и длине интрона; 3) все, что можно получить из «ACCESSION M84466», откуда мы сей интрон почерпнули, а там можно найти ссылки, сорт растения, экотип и т.д.; 4) природный кодоновый (аминокислотный) контекст:

Экзон 5’ Экзон 3’

AACTTATTAG|GTTCTTGAAT

N L L G S *

T Y * V L E

L I R F L N

Например, в этом случае в природной мРНК – контекст **LI{R}FLN,** т.е. кодируемый триплет аргинина **AGG** «разрывается» **AG|G.** Природный контекст важнее, поскольку в «искусственном» (два других могли бы иметь право на существование) есть потенциальная проблема с «codon pair bias».

Это все о программе А.

Программа Б:

1. Ограничим кодоновый контекст «табачной» (любой) базы, кодируемыми аминокислотами, возьмем только по одной аминокислоте (это число может меняться) «справа» и «слева» от центральной: **I{R}F**.
2. Найдем в заданной аминокислотной последовательности (например, в репликазе вируса желтухи свеклы) эти три аминокислоты:

MAFLNVSAVPSCAFAPAFAPHAGASPIVPDSFPCVPRYSDDISHFRLTLSLDFSVPRPLSLNARVHLSASTDNPLPSLPLGFHAETFVLELNGSSAP
FSIPSRHIDFVVNRPFSVFPTEVL(1)SVSSLRTPSRLFALLCDFFLYCSKPGPCVEIASFSTPPPCLVSNCVAQIPTHAEMESIRF(1)PTKTLPA
GRFLQFHKRKYTKRPETLIIHESGLALKTSALGVTSKPNSRPITVKSASGEKYEAYEISRKDFERSRRRQQTPRVRSHKPRKINKAVEPFFFPEEPK
KDKRKRASLPTEDEGFITFGTLRFPLSETPKEEPRLPKFREVEIPVVKKHAVPAVVSKPVRTFRPVATTGAEYVNARNQCSRRPRNHPILRSASYTF
GFKKMPLQRFMKEKKEYYVKRS(1)KVVSSCSVTKSPLEALASILKNLPQYSYNSERLKFYDHFIGDDFEIEVHPLRGGKLSVLLILPKGEAYCVVT
AATPQYHAALTIARGDRPRVGELLQYRPGEGLCYLAHAALCCALQKRTFREEDFFVGMYPTKFVFAKRLTEKLGPSALKHPVRGRQVSRSLFHCDVA
SAFSSPFYSLPRFIGGVEEEAPEITSSLKHKAIESVYERVSIHKDNLLARSVEKDLIDFKDEIKSLSKEKRS(2)VTVPFYMGEAVQSGLTRAYPQF
NLSFTHSVYSDHPAAAGSRLLENETLASMAKSSFSDIGGCPLFHIKRGSTDYHVCRPIYDMKDAQRRVSRELQARGL(1)VENLSREQLVEAQARVS
VCPHTLGNCNVKSDVLIMVQVYDASLNEIASAMVLKESKVA(1)YLTMVTPGELLDEREAFAIDALGCDVVVDTRRDMVQYKFGSSCYCHKLSNIKS
IMLTPAFTFSGNLFSVEMYENRMGVNYYKITRSAYSPEIRGVKTLRYRRACTEVVQVKLPRFDKTLKTFLSGYDYIYLDAKFVSRVFDYVVSNCSVV
NSKTFEWVWSYIKSSKSRVVISGKVIHRDVHIDLKHSECFAAVMLAVGVRSRTTTEFLAKNLNYYTGDASCFETIRF(2)LFREWSRRAYAEINRSF
RKLMKSILSAGLDYEFLDLDNSLQHLLEYSEVEVRVSIAQNGEVDCNEENRVLTEIIAEAADRKSIAQGLSGALSSVPTQPRGGLRGGSRRSGV(1)
SFLYNLVEEVGNLFFSVGDAVRFLVKVFKTFSDSPIFRVVRMFLDLAEAASPFVSVVSLCAWLREAVSAFSSWVADRTVSESVKTFVNRTVKRFLNF
MSAKTLTKKFFRFFLSASALAKTVVRKAKVILEAYWEVWFESILSDSGEYSAVEFCSSVVITLLTNSGRLLPGFSPSAIITEVL(2)LDLATKISIE
VL(3)LKQISPADSTASSALYRRVLSEILSNFRTMGEHGIFTKVFLLCGFLPVFVRKCVALCVPGDMATYARFLEYGVDDLFFLGRSVNSIKNYLCV
VAAGLVDSIVDSVVLKLSGV(2)AKERVLGFKSKIIKNFLNVFRKAKVVTRTSSSTDLSEDEYFSCDESKPGLRGGSSRFTLSRLLDIFFNFLKSSK
LVIENACFSAYERIERNRMKLYFFPLNSSEEEARRLIRCAGDFDYLSDSAFDEDEMLRQAFEQYYSSDDESVTYDGKPTVLRSYLNVSRRFLETFCNG
PKFFVKVSNYFKALYSRLLRVLPWVDRNLSDSPGLKGGNEKALLAKFFKTCVITACECVSQICCLRLIRLCWGTPACGLVRLFYITYSSTRVLSRVV
VAVAVCPLLVRNELDGLSDGLTNMGVSVFRRLFVALRRALSAYSNSALRRKIIEFIFGNIHHPFDVAVIETNEVAPEPLSPEVDIDVDCDFGSDSES
VSSDEVASNPRPGLHGGSRRSSNFLTSLVKVVFKLARRIPRLLFRLRNFVAYFVERRLASKRLKTFIGLARLFDNFSLTSVVYLLQEYDSVLNAFID
VELILLNSGSVNVLPLVSWVRGSLTKLAEAIVGSGFASFLGRMCCRVSDWCSSSSNAGCNFMSPVRTKGKFVPPSSSGSTASMYERLEALESDIREH
VLSTCRVGSDEEEERPKEVTEPGIEHTSEDVVPIRSHSQPLSGGECSYSEDREENERANLLPHVSKIVSERRGL(2)ETARRNKRTLHGVSEFLNAI
NTSNEQPRPIIVDHSPESRALTNSVREFYYLQELALFELSCKLREYYDQLKVA(2)NFNRQECLCDKDEDMFVLRAGQGVVSGRNSRLPLKHFKGHE
FCFRSGGLVPYDGTSRVDTIFHTQTNFVSANALLSGYLSYRTFTFTNLSANVLLYEAPPGGGKTTTLIKVFCETFSKVNSLILTANKSSREEILAKV
NRIVLDEGDTPLQTRDRILTIDSYLMNNRGL(3)TCKVLYLDECFMVHAGAAVACIEFTKCDSAILFGDSRQIRYGRCSELDTAVLSDLNRFVDDES
RVYGEVSYRCPWDVCAWLSTFYPKTVATTNLVSAGQSSMQVREIESVDDVEYSSEFVYLTMLQSEKKDLLKSFGKRS(3)RSSVEKPTVLTVHEAQG
E(1)TYRKVNLVRTKFQEDDPFRSENHITVALSRHVESLTYSVLSSKRDDAIAQAIVKAKQLVDAYRVYPTSFGGSTLDVSVNPSTSDRSKCKASSA
PYEVINSFLESVVPGTTSVDFGDVSEEMGTQVFESGADNVVIRDSAPVNKSTDHDPQRV*

Есть такие! Встречаются дважды (показано красным на желтом фоне)! Остальные «подчеркивания» - «интроны» арабидопсиса.

3. Для сканирования заданной аминокислотной последовательности на предмет наличия «аминокислотного» контекста интрона из заданной базы (или баз), неплохо задавать предельный размер потенциального интрона.

Так работаю я, а должна работать программа Б.

4. А и Б могут быть совмещены в одном интерфейсе.