Open for Innovation
KNIME

KNIME Pros Learnathon

Building Reliable and Reusable Components

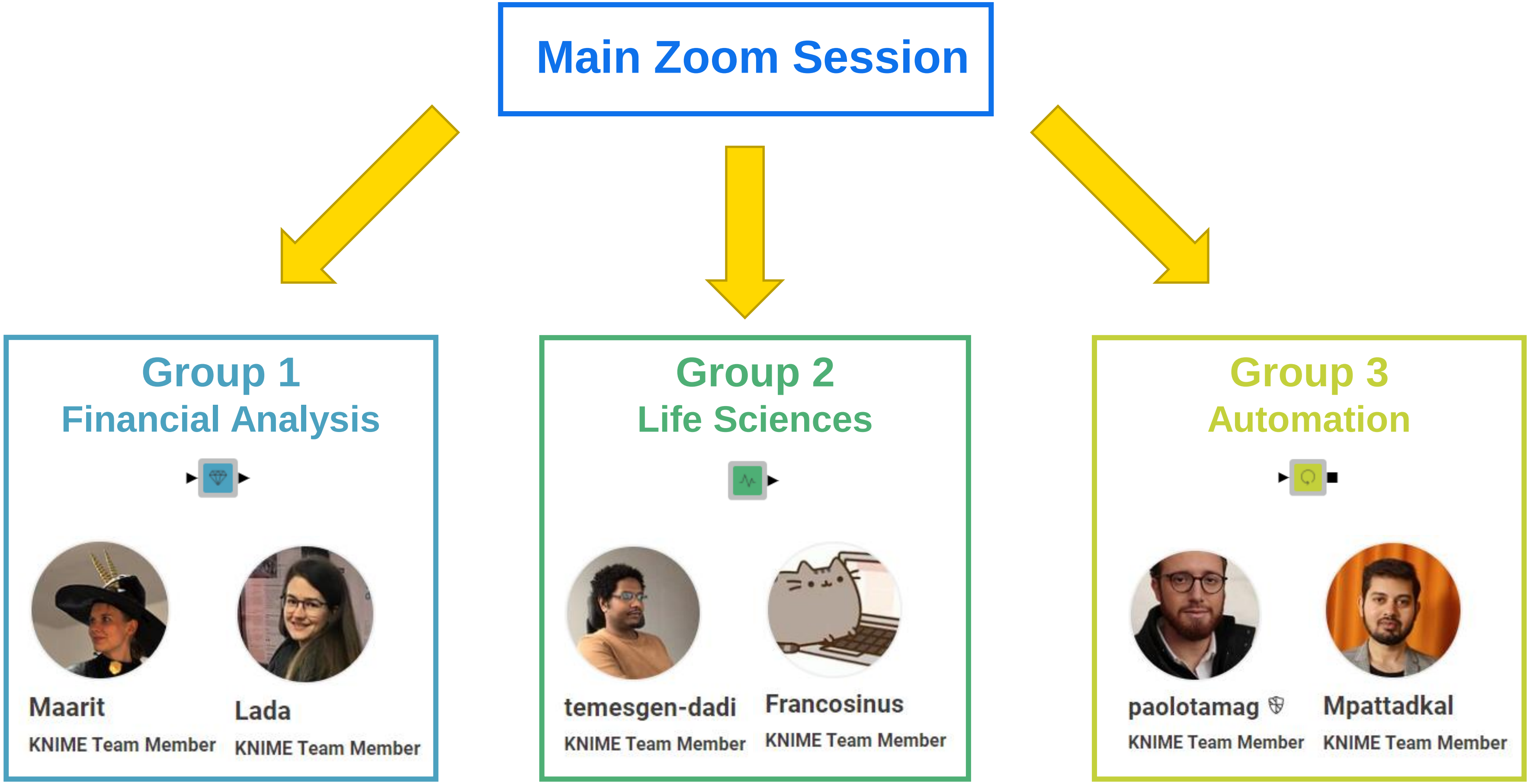
Group 2: Life Sciences

KNIME Team

@KNIME
#Learnathon

tinyurl.com/KNIME-Pros-Stuff

See you soon in a Breakout Room!



Main Zoom Session

Group 1
Financial Analysis

Maarit
KNIME Team Member

Lada
KNIME Team Member

Group 2
Life Sciences

temesgen-dadi
KNIME Team Member

Francosinus
KNIME Team Member

Group 3
Automation

paolotamag
KNIME Team Member

Mpattadkal
KNIME Team Member

© 2020 KNIME AG. All rights reserved.

tinyurl.com/KNIME-Pros-Stuff

Three Parallel Tracks via Zoom Breakout Rooms!

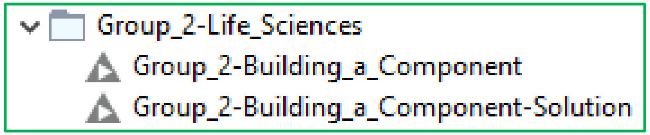
Main Zoom Session

Welcome to:
Group 2.
Life Sciences

Group 2
Life Sciences

temesgen-dadi
KNIME Team Member

Francosinus
KNIME Team Member



Group_2-Life_Sciences
- Group_2-Building_a_Component
- Group_2-Building_a_Component-Solution

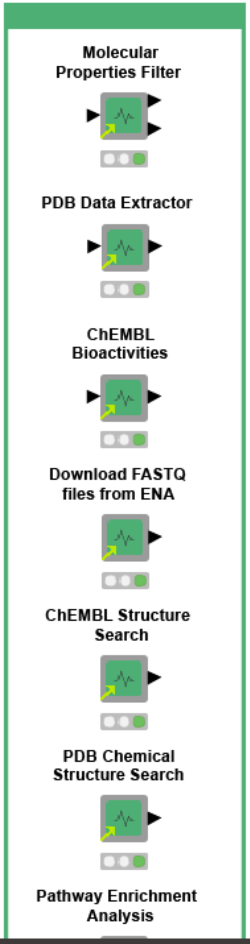
- Download exercises from:
  tinyurl.com/KNIME-Pros-Stuff

OR

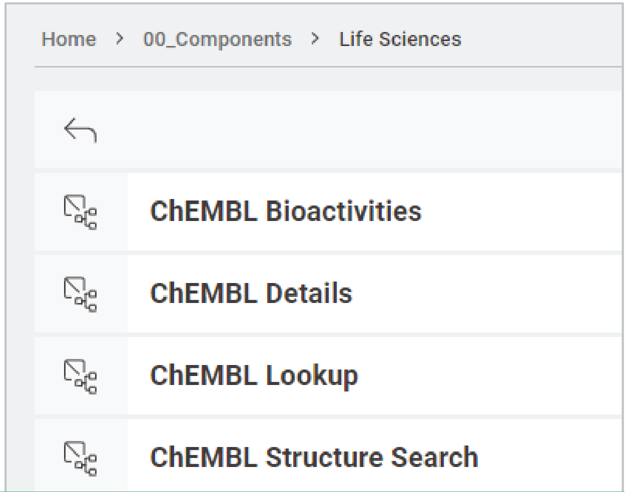
- Search **hub.knime.com** :
  “KNIME Pros Learnathon Group 2”

© 2020 KNIME AG. All rights reserved.

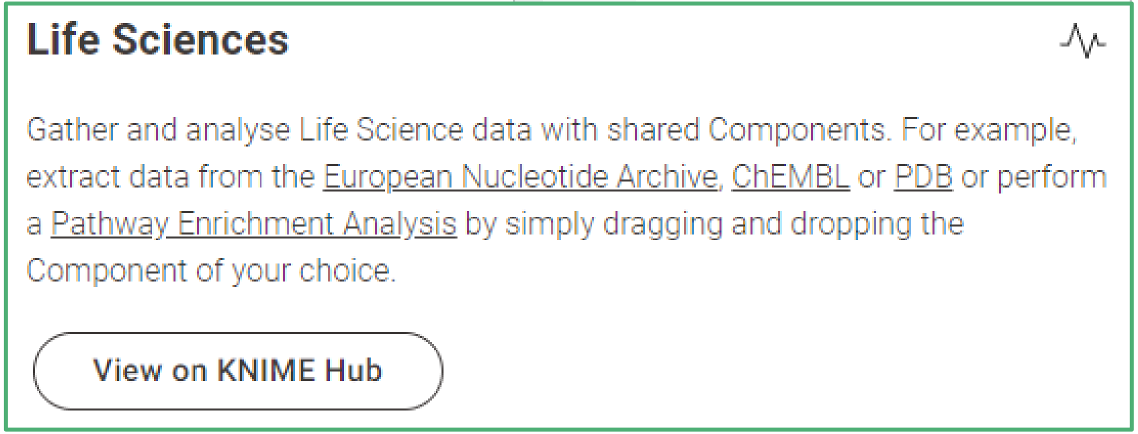
Life Sciences Verified Components


Molecular Properties Filter
PDB Data Extractor
ChEMBL Bioactivities
Download FASTQ files from ENA
ChEMBL Structure Search
PDB Chemical Structure Search
Pathway Enrichment Analysis

Available on hub.knime.com



- ChEMBL Bioactivities
- ChEMBL Details
- ChEMBL Lookup
- ChEMBL Structure Search

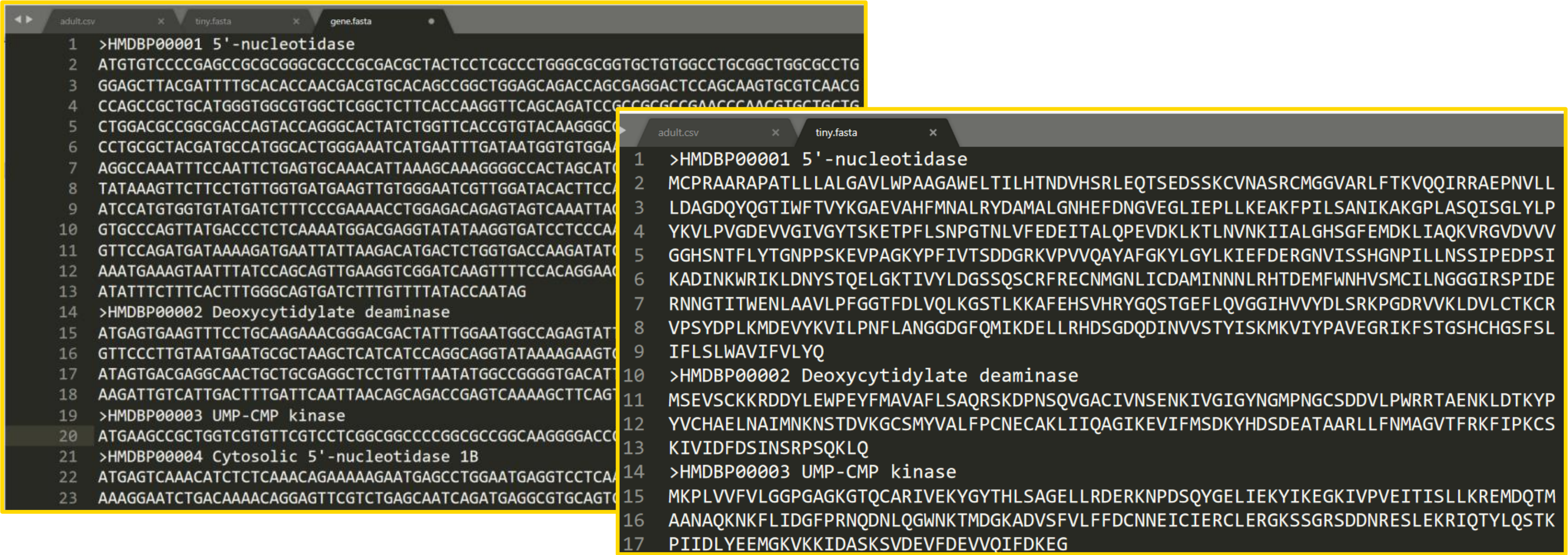


Life Sciences

Gather and analyse Life Science data with shared Components. For example, extract data from the European Nucleotide Archive, ChEMBL or PDB or perform a Pathway Enrichment Analysis by simply dragging and dropping the Component of your choice.

FASTA File Format

- common file format in bioinformatics and biochemistry
- text-based format for representing either nucleotide sequences or amino acid (protein) sequences
- allows for sequence names and comments to precede the sequences



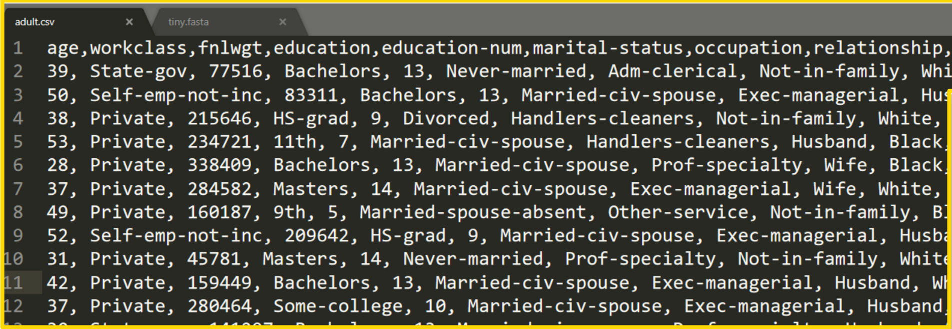
```
>HMDBP00001 5'-nucleotidase
ATGTGTCCCCGAGCCGCGCGGGCGCCCGCGACGCTACTCCTCGCCCTGGGCGCGGTGCTGTGGCCTGCGGCTGGCGCCTG
GGAGCTTACGATTTTGCACACCAACGACGTGCACAGCCGGCTGGAGCAGACCAGCGAGGACTCCAGCAAGTGCGTCAACG
CCAGCCGCTGCATGGGTGGCGTGGCTCGGCTCTTCACCAAGGTTCAGCAGATCCGCCGCGCCGAACCCAACGTGCTGCTG
CTGGACGCCGGCGACCAGTACCAGGGCACTATCTGGTTCACCGTGTACAAGGGCC
CCTGCGCTACGATGCCATGGCACTGGGAAATCATGAATTTGATAATGGTGTGGAA
AGGCCAAATTTCCAATTCTGAGTGCAAACATTAAAGCAAAGGGGCCACTAGCATC
TATAAAGTTCTTCCTGTTGGTGATGAAGTTGTGGGAATCGTTGGATACACTTCCA
ATCCATGTGGTGTATGATCTTTCCCGAAAACCTGGAGACAGAGTAGTCAAATTAC
GTGCCCAGTTATGACCCTCTCAAAATGGACGAGGTATATAAGGTGATCCTCCCAA
GTTCCAGATGATAAAAGATGAATTATTAAGACATGACTCTGGTGACCAAGATATC
AAATGAAAGTAATTTATCCAGCAGTTGAAGGTCGGATCAAGTTTTCCACAGGAAC
ATATTTCTTTCACTTTGGGCAGTGATCTTTGTTTTATACCAATAG
>HMDBP00002 Deoxycytidylate deaminase
ATGAGTGAAGTTTCCTGCAAGAAACGGGACGACTATTTGGAATGGCCAGAGTATT
GTTCCCTTGTAATGAATGCGCTAAGCTCATCATCCAGGCAGGTATAAAAGAAGTC
ATAGTGACGAGGCAACTGCTGCGAGGCTCCTGTTTAATATGGCCGGGGTGACATT
AAGATTGTCATTGACTTTGATTCAATTAACAGCAGACCGAGTCAAAAGCTTCAGT
>HMDBP00003 UMP-CMP kinase
ATGAAGCCGCTGGTCGTGTTCGTCCTCGGCGGCCCCGGCGCCGGCAAGGGGACCC
>HMDBP00004 Cytosolic 5'-nucleotidase 1B
ATGAGTCAAACATCTCTCAAACAGAAAAAGAATGAGCCTGGAATGAGGTCCTCAA
AAAGGAATCTGACAAAACAGGAGTTCGTCTGAGCAATCAGATGAGGCGTGCAGTC
```

```
>HMDBP00001 5'-nucleotidase
MCPRAARAPATLLLALGAVLWPAAGAWELTILHTNDVHSRLEQTSEDSSKCVNASRCMGGVARLFTKVQQIRRAEPNVLL
LDAGDQYQGTIWFTVYKGAEVAHFMNALRYDAMALGNHEFDNGVEGLIEPLLKEAKFPILSANIKAKGPLASQISGLYLP
YKVLPVGDEVVGIVGYTSKETPFLSNPGTNLVFEDEITALQPEVDKLKTLNVNKIIALGHSGFEMDKLIAQKVRGVDVVV
GGHSNTFLYTGNPPSKEVPAGKYPFIVTSDDGRKVPVVQAYAFGKYLGYLKIEFDERGNVISSHGNPILLNSSIPEDPSI
KADINKWRIKLDNYSTQELGKTIVYLDGSSQSCRFRECNMGNLICDAMINNNLRHTDEMFWNHVSMCILNGGGIRSPIDE
RNNGTITWENLAAVLPFGGTFDLVQLKGSTLKKAFEHSVHRYGQSTGEFLQVGGIHVVYDLSRKPGDRVVKLDVLCTKCR
VPSYDPLKMDEVYKVILPNFLANGGDGFQMIKDELLRHDSGDQDINVVSTYISKMKVIYPAVEGRIKFSTGSHCHGSFSL
IFLSLWAVIFVLYQ
>HMDBP00002 Deoxycytidylate deaminase
MSEVSCKKRDDYLEWPEYFMAVAFLSAQRSKDPNSQVGACIVNSENKIVGIGYNGMPNGCSDDVLPWRRTAENKLDTKYP
YVCHAELNAIMNKNSTDVKGCSMYVALFPCNECAKLIIQAGIKEVIFMSDKYHDSDEATAARLLFNMAGVTFRKFIPKCS
KIVIDFDSINSRPSQKLQ
>HMDBP00003 UMP-CMP kinase
MKPLVVFVLGGPGAGKGTQCARIVEKYGYTHLSAGELLRDERKNPDSQYGELIEKYIKEGKIVPVEITISLLKREMDQTM
AANAQKNKFLIDGFPRNQDNLQGWNKTMDGKADVSFVLFFDCNNEICIERCLERGKSSGRSDDNRESLEKRIQTYLQSTK
PIIDLYEEMGKVKKIDASKSVDEVFDEVVQIFDKEG
```

© 2020 KNIME AG. All rights reserved.

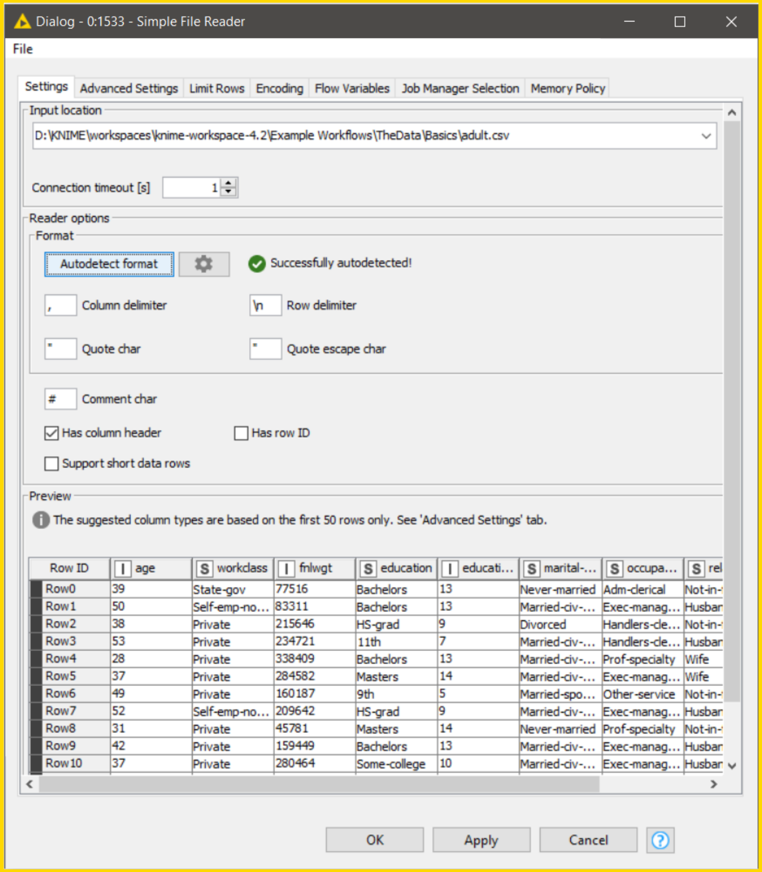
tinyurl.com/KNIME-Pros-Stuff

KNIME Nodes (File Readers)



adult.csv tiny.fasta

```
age,workclass,fnlwgt,education,education-num,marital-status,occupation,relationship,race,sex,capital-gain,capital-loss,hours-per-week,native-country,income
39, State-gov, 77516, Bachelors, 13, Never-married, Adm-clerical, Not-in-family, White, Male, 2174, 0, 40, United-States, <=50K
50, Self-emp-not-inc, 83311, Bachelors, 13, Married-civ-spouse, Exec-managerial, Hus
38, Private, 215646, HS-grad, 9, Divorced, Handlers-cleaners, Not-in-family, White,
53, Private, 234721, 11th, 7, Married-civ-spouse, Handlers-cleaners, Husband, Black
28, Private, 338409, Bachelors, 13, Married-civ-spouse, Prof-specialty, Wife, Black
37, Private, 284582, Masters, 14, Married-civ-spouse, Exec-managerial, Wife, White,
49, Private, 160187, 9th, 5, Married-spouse-absent, Other-service, Not-in-family, Bl
52, Self-emp-not-inc, 209642, HS-grad, 9, Married-civ-spouse, Exec-managerial, Husba
31, Private, 45781, Masters, 14, Never-married, Prof-specialty, Not-in-family, White
42, Private, 159449, Bachelors, 13, Married-civ-spouse, Exec-managerial, Husband, Wh
37, Private, 280464, Some-college, 10, Married-civ-spouse, Exec-managerial, Husband
```



Dialog - 0:1533 - Simple File Reader

File

Settings | Advanced Settings | Limit Rows | Encoding | Flow Variables | Job Manager Selection | Memory Policy

Input location

D:\KNIME\workspaces\knime-workspace-4.2\Example Workflows\TheData\Basics\adult.csv

Connection timeout [s] 1

Reader options

Format

Autodetect format ✓ Successfully autodetected!

Column delimiter: ,

Row delimiter: \n

Quote char: "

Quote escape char: "

Comment char: #

☑ Has column header ☐ Has row ID

☐ Support short data rows

Preview

The suggested column types are based on the first 50 rows only. See 'Advanced Settings' tab.

Row ID	I age	S workclass	I fnlwgt	S education	I educati...	S marital-...	S occupa...	S rel
Row0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-
Row1	50	Self-emp-no...	83311	Bachelors	13	Married-civ-...	Exec-manag...	Husbar
Row2	38	Private	215646	HS-grad	9	Divorced	Handlers-cle...	Not-in-
Row3	53	Private	234721	11th	7	Married-civ-...	Handlers-cle...	Husbar
Row4	28	Private	338409	Bachelors	13	Married-civ-...	Prof-specialty	Wife
Row5	37	Private	284582	Masters	14	Married-civ-...	Exec-manag...	Wife
Row6	49	Private	160187	9th	5	Married-spo...	Other-service	Not-in-
Row7	52	Self-emp-no...	209642	HS-grad	9	Married-civ-...	Exec-manag...	Husbar
Row8	31	Private	45781	Masters	14	Never-married	Prof-specialty	Not-in-
Row9	42	Private	159449	Bachelors	13	Married-civ-...	Exec-manag...	Husbar
Row10	37	Private	280464	Some-college	10	Married-civ-...	Exec-manag...	Husbar

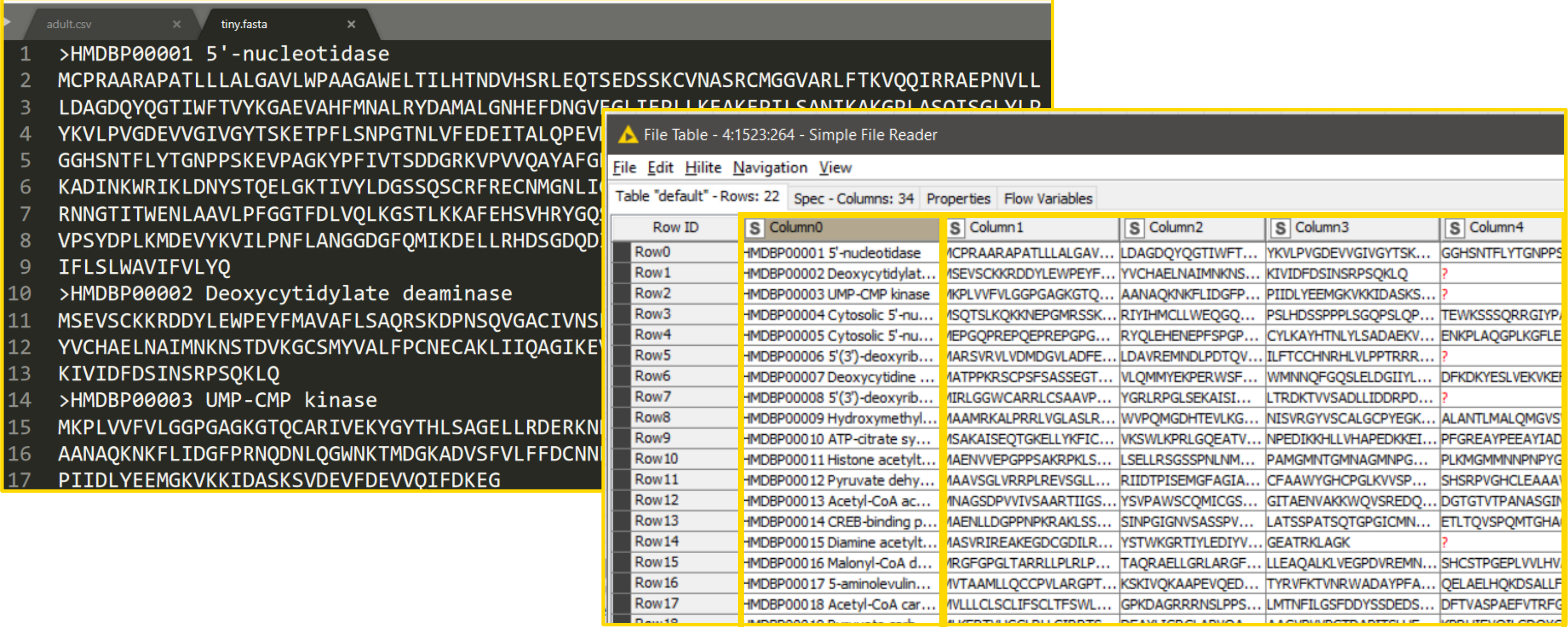
OK
Apply
Cancel

File Reader **CSV Reader** **CSV Reader (Labs)** **Simple File Reader**

- Well suited to read tabular data

tinyurl.com/KNIME-Pros-Stuff

FASTA File as Tabular Data


adult.csv
tiny.fasta
File Table - 4:1523:264 - Simple File Reader
File Edit Hilite Navigation View
Table "default" - Rows: 22
Spec - Columns: 34
Properties
Flow Variables

```
>HMDBP00001 5'-nucleotidase
MCPRAARAPATLLLALGAVLWPAAGAWELTILHTNDVHSRLEQTSEDSSKCVNASRCMGGVARLFTKVQQIRRAEPNVLL
LDAGDQYQGTIWFTVYKGAEVAHFMNALRYDAMALGNHEFDNGVEGLIEPLLKEAKFPILSANIKAKGPLASQISGLYLP
YKVLPVGDEVVGIVGYTSKETPFLSNPGTNLVFEDEITALQPEV
GGHSNTFLYTGNPPSKEVPAGKYPFIVTSDDGRKVPVVQAYAFG
KADINKWRIKLDNYSTQELGKTIVYLDGSSQSCRFRECNMGNLI
RNNGTITWENLAAVLPFGGTFDLVQLKGSTLKKAFEHSVHRYGQ
VPSYDPLKMDEVYKVILPNFLANGGDGFQMIKDELLRHDSGDQD
IFLSLWAVIFVLYQ
>HMDBP00002 Deoxycytidylate deaminase
MSEVSCKKRDDYLEWPEYFMAVAFLSAQRSKDPNSQVGACIVNS
YVCHAELNAIMNKNSTDVKGCSMYVALFPCNECAKLIIQAGIKE
KIVIDFDSINSRPSQKLQ
>HMDBP00003 UMP-CMP kinase
MKPLVVFVLGGPGAGKGTQCARIVEKYGYTHLSAGELLRDERKN
AANAQKNKFLIDGFPRNQDNLQGWNKTMDGKADVSFVLFFDCNN
PIIDLYEEMGKVKKIDASKSVDEVFDEVVQIFDKEG
```

Row ID	S Column0	S Column1	S Column2	S Column3	S Column4
Row0	HMDBP00001 5'-nucleotidase	MCPRAARAPATLLLALGAV...	LDAGDQYQGTIWFT...	YKVLPVGDEVVGIVGYTSK...	GGHSNTFLYTGNPPS
Row1	HMDBP00002 Deoxycytidylat...	MSEVSCKKRDDYLEWPEYF...	YVCHAELNAIMNKNS...	KIVIDFDSINSRPSQKLQ	?
Row2	HMDBP00003 UMP-CMP kinase	MKPLVVFVLGGPGAGKGTQ...	AANAQKNKFLIDGFP...	PIIDLYEEMGKVKKIDASKS...	?
Row3	HMDBP00004 Cytosolic 5'-nu...	MSQTSLKQKKNEPGMRSSK...	RIYIHMCLLWEQGQ...	PSLHDSSPPPLSGQPSLQP...	TEWKSSSQRRGIYPA
Row4	HMDBP00005 Cytosolic 5'-nu...	MEPGQPREPQEPREPGPG...	RYQLEHENEPFSPGP...	CYLKAYHTNLYLSADAEKV...	ENKPLAQGPLKGFLE
Row5	HMDBP00006 5'(3')-deoxyrib...	MARSVRVLVDMDGVLADFE...	LDAVREMNDLPDTQV...	ILFTCCHNRHLVLPPTRRR...	?
Row6	HMDBP00007 Deoxycytidine ...	MATPPKRSCPSFSASSEGT...	VLQMMYEKPERWSF...	WMNNQFGQSLELDGIIYL...	DFKDKYESLVEKVKEI
Row7	HMDBP00008 5'(3')-deoxyrib...	MIRLGGWCARRLCSAAVP...	YGRLRPGLSEKAISI...	LTRDKTVVSADLLIDDRPD...	?
Row8	HMDBP00009 Hydroxymethyl...	MAAMRKALPRRLVGLASLR...	WVPQMGDHTEVLKG...	NISVRGYVSCALGCPYEGK...	ALANTLMALQMGVSV
Row9	HMDBP00010 ATP-citrate sy...	MSAKAISEQTGKELLYKFIC...	VKSWLKPRLGQEATV...	NPEDIKKHLLVHAPEDKKEI...	PFGREAYPEEAYIAD
Row10	HMDBP00011 Histone acetylt...	MAENVVEPGPPSAKRPKLS...	LSELLRSGSSPNLNM...	PAMGMNTGMNAGMNPG...	PLKMGMMNNPNPYG
Row11	HMDBP00012 Pyruvate dehy...	MAAVSGLVRRPLREVSGLL...	RIIDTPISEMGFAGIA...	CFAAWYGHCPGLKVVSP...	SHSRPVGHCLEAAAV
Row12	HMDBP00013 Acetyl-CoA ac...	MNAGSDPVVIVSAARTIIGS...	YSVPAWSCQMICGS...	GITAENVAKKWQVSREDQ...	DGTGTVTPANASGIN
Row13	HMDBP00014 CREB-binding p...	MAENLLDGPPNPKRAKLSS...	SINPGIGNVSASSPV...	LATSSPATSQTGPGICMN...	ETLTQVSPQMTGHAG
Row14	HMDBP00015 Diamine acetylt...	MASVRIREAKEGDCGDILR...	YSTWKGRTIYLEDIYV...	GEATRKLAGK	?
Row15	HMDBP00016 Malonyl-CoA d...	MRGFGPGLTARRLLPLRLP...	TAQRAELLGRLARGF...	LLEAQALKLVEGPDVREMN...	SHCSTPGEPLVVLHVA
Row16	HMDBP00017 5-aminolevulin...	MVTAAMLLQCCPVLARGPT...	KSKIVQKAAPEVQED...	TYRVFKTVNRWADAYPFA...	QELAELHQKDSALLF
Row17	HMDBP00018 Acetyl-CoA car...	MVLLLCLSCLIFSCLTFSWL...	GPKDAGRRRNSLPPS...	LMTNFILGSFDDYSSDEDS...	DFTVASPAEFVTRFG

SequenceID

← Sequence →

- Re-think Row/Column delimiters
  - “>” Row delimitr
  - “\n” Column delimeter

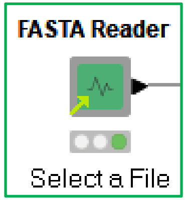
Today’s Challenge: FASTA Reader Component

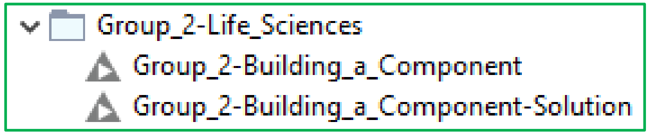
- Download Exercises from **link** and import .knar file to your KNIME Analytics Platform LOCAL Workspace!

OR

- Download from KNIME Hub!

hub.knime.com/knime/spaces/Education/latest/Learnathons
or Search for “**KNIME Pros Learnathon - Group 2**”




Group_2-Life_Sciences
- Group_2-Building_a_Component
- Group_2-Building_a_Component-Solution

tinyurl.com/KNIME-Pros-Stuff

© 2020 KNIME AG. All rights reserved.