Hasso Plattner Institut
Digital Engineering · Universität Potsdam

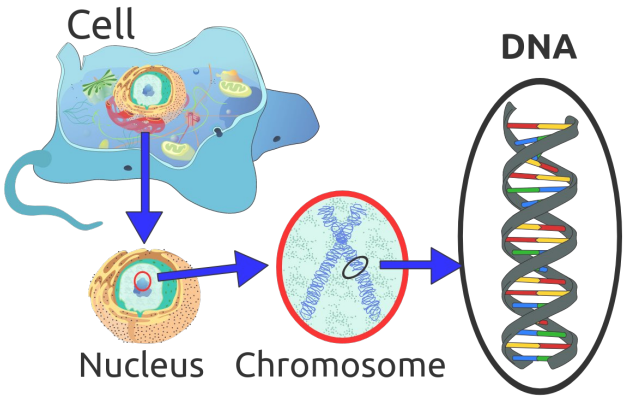
Integrative Gene Selection

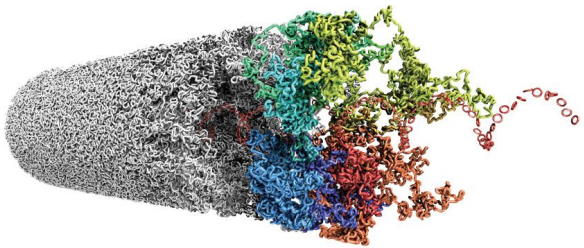
-

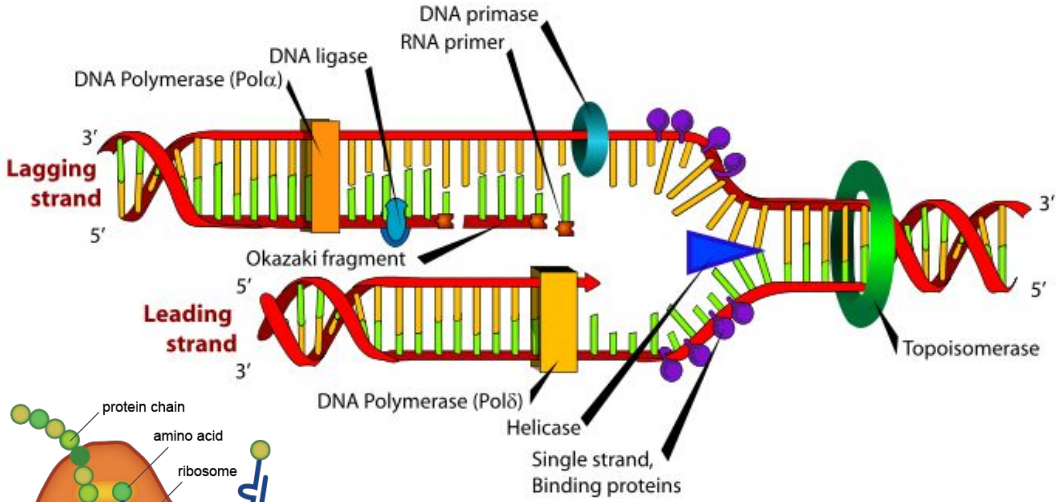
Intermediate Presentation

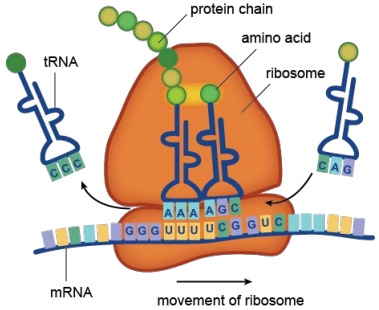
Trends in Bioinformatics WS 2018/2019

Biology Recap





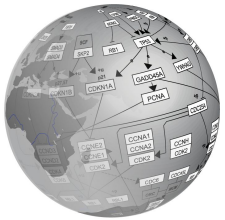




Biology Recap

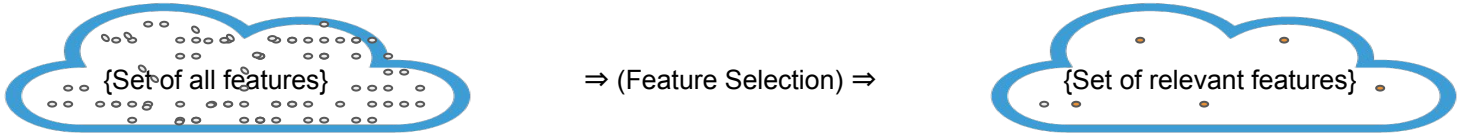
		ENSG00000000003	ENSG00000000005	ENSG00000000419	ENSG00000000457	ENSG00000000460	ENSG00000000938	ENSG00000000971	ENSG00000001036	ENSG00000001084	ENSG00000000011
2	TCGA-19-1389-02A-21R-2005-01	11.7792217414034	1.89364932890714	10.2984404906165	8.82944103729988	8.0542873101456	9.947585043288 54	12.4610251009447	12.6365488867816	12.803213351306	10.3150069248
3	TCGA-32-1970-01A-01R-1850-01	11.4920850420329	4.32278842625045	9.39890785671076	8.90487668099722	8.18094740444371	8.71004929184007	10.4209629013198	10.9260948180845	11.3169553347747	10.2477539598
4	TCGA-28-2499-01A-01R-1850-01	13.3325817590725	3.65988011150083	10.1069561457736	8.74359115369343	7.61083922743939	8.53535837188253	11.9563846434749	10.9723865561388	10.4630402470726	10.1128316168
5	TCGA-06-2567-01A-01R-1849-01	11.5232355093901	4.00039138581705	9.76192997765078	9.18009034939688	8.9109653881182	9.32100723890391	10.4379486546418	10.9770972752061	12.0447760154288	10.5835808316
6	TCGA-26-5139-01A-01R-1850-01	12.2141481981795	3.67053989301986	9.93385952115899	8.98163879119303	8.01269761983059	8.84233309502494	11.3157834789084	11.5519669726633	10.5810982537438	10.0784267484
7	TCGA-14-0790-01B-01R-1849-01	12.1619979609256	4.26566921129679	9.97351631059356	9.02106440617208	8.18794612594937	7.59938650805111	11.1437216091891	11.747476675689	10.9195881621739	9.97889674261
8	TCGA-06-0174-01A-01R-1849-01	11.8961825784019	2.87673316311828	9.83582262314522	9.05258559063893	8.85723274269382	8.59152679065233	8.72591359900766	11.0551669309406	10.7153893942884	10.6109144250
9	TCGA-06-0744-01A-01R-1849-01	11.8977228680519	3.76378477187468	10.2513534186818	9.16869334984741	8.10956423114048	8.14196665205784	9.68601696129216	11.3687934020961	10.4884777575476	9.47696893807
10	TCGA-41-4097-01A-01R-1850-01	11.283933012877	2.744003592804 34	9.79499529398278	8.99714844764523	7.68189470776302	9.02685347302661	12.2075318446436	11.1729068849665	10.589430549805	9.64024865595
11	TCGA-32-2638-01A-01R-1850-01	12.3434299411025	5.021722151115	10.2511334569768	9.06309749746212	8.19216225467997	8.6923878902482	9.47267965789312	11.2082062667692	10.1401841574084	10.0045945474
12	TCGA-27-2526-01A-01R-1850-01	12.7165302680384	5.5519104039832	10.9548088449485	8.69434473542891	8.34188648944638	9.7227677123402	11.3728317435249	11.542926370051	11.4475267478933	9.52331756024
13	TCGA-06-5410-01A-01R-1849-01	10.9156392211967	3.0651101506293	10.3627189353996	8.96299155375621	8.48208545889823	11.0000826476833	11.6922256881476	11.7109059341728	11.0935868695773	10.0161494654
14	TCGA-27-1835-01A-01R-1850-01	11.6317701680024	3.43273715625251	10.3011454608069	9.26125055665528	8.0635316206427	8.18366869381058	11.6480459274422	11.0918711638985	10.9975016467621	10.1066739531
15	TCGA-06-0686-01A-01R-1849-01	11.4995199858581	2.886737634139 31	9.74644350241758	8.87308444433284	8.56660091710499	9.58382031821748	9.28243374725845	10.889731349267	10.8473397287558	10.3607929355
16	TCGA-16-1045-01B-01R-1850-01	11.8719641187614	4.82572639688055	10.0189715455784	9.0446915618307	8.07070572557224	10.2012811178975	12.8050175146146	10.8178556901018	11.2729614600611	9.99845303590
17	TCGA-02-0047-01A-01R-1849-01	11.4420485202034	3.23183810777129	9.5856215654832	8.91370793270221	8.15124733777834	9.40990014810783	9.53756096918312	11.2855214070653	11.2425625974807	10.3674576733
18	TCGA-15-1444-01A-02R-1850-01	12.0189246051033	4.10218992965831	9.75483916402011	8.95951704267504	8.58522184941783	9.76736456842158	9.40811038383531	10.4807777124695	11.9219840517328	10.3352535023
19	TCGA-19-2629-01A-01R-1850-01	11.2997177038121	2.99119545357813	9.66023816460349	8.90375915434759	8.32983958299574	8.52804943302458	10.3929896745369	9.13749577729725	10.9855018965556	10.0462066826
20	TCGA-06-0178-01A-01R-1849-01	12.7721846840402	3.84191347169612	10.9019716222042	10.0455627203785	9.94930350185811	7.96612265305727	8.39511174555155	10.1690558025555	10.8577831489131	10.0473523424
21	TCGA-28-1747-01C-01R-1850-01	12.1741625460043	3.77237183906278	9.86118120416317	9.01448903689638	8.30856055148352	9.87237389128326	9.88624372929727	10.6889479952912	11.7104707533006	9.78475199044
22	TCGA-14-2554-01A-01R-1850-01	14.2459146539805	8.13808728123304	10.7859422702553	9.06348586274211	8.50823908360417	9.00315554819953	11.3491456859856	12.0314882848862	10.8714464650062	9.76025639216
23	TCGA-06-0680-11A-32R-A36H-07	9.36794940464877	3.43293476160042	9.83384658594166	9.47921321737844	6.64715468895986	7.98672518823651	9.63572602278761	9.66409821174503	11.0828789538052	10.4237130537
24	TCGA-28-1753-01A-01R-1850-01	11.2623632817346	2.89045187431281	9.88127529414005	8.86138358968327	8.21497349543574	10.9555359030166	11.7621125079785	11.2149860468849	11.0464632788109	10.4482329755
25	TCGA-76-4929-01A-01R-1850-01	12.4627013621312	6.59563960154395	10.1843965827281	9.35349199366977	7.44674031087527	9.75593090544685	9.74685914547499	10.2295981362216	11.0126192281517	11.0875904575
26	TCGA-19-4065-02A-11R-2005-01	11.4741699006582	3.17712120559885	10.1626214403749	8.92710312722997	7.9115643305666	9.41248231753299	12.1973562185909	12.0978686433899	10.4023969189665	10.2333195338
27	TCGA-32-2616-01A-01R-1850-01	11.3905867648387	4.59055211852038	10.0143149780262	8.65273929682339	8.34093171932781	8.94629420057476	7.47877575845708	11.8962059858922	10.600455240484	10.2183382151

Example Data from TCGA: 2GB csv, few samples but thousands of columns!



WIKIPATHWAYS
Pathways for the People

Feature Selection



- **Short, fat data (p >> n)**

  RNAseq data typically covers p-thousands of genes but only few n-hundred samples

- **Incomplete view on cell process**

  RNAseq data covers a complete snapshot of the gene activity of a cell; only a single point in time - not more, not less...

- **Gene interactions**

  Genes react to changes of partners in the interaction network, interactions have to be taken into account

- **Biological Relevance**

  Statistical approaches only concentrate on statistical relevance, ignoring the biological relevance of genes, like disease driver genes, while only showing the respective behaviour of increased gene expression profiles of affected pathway genes

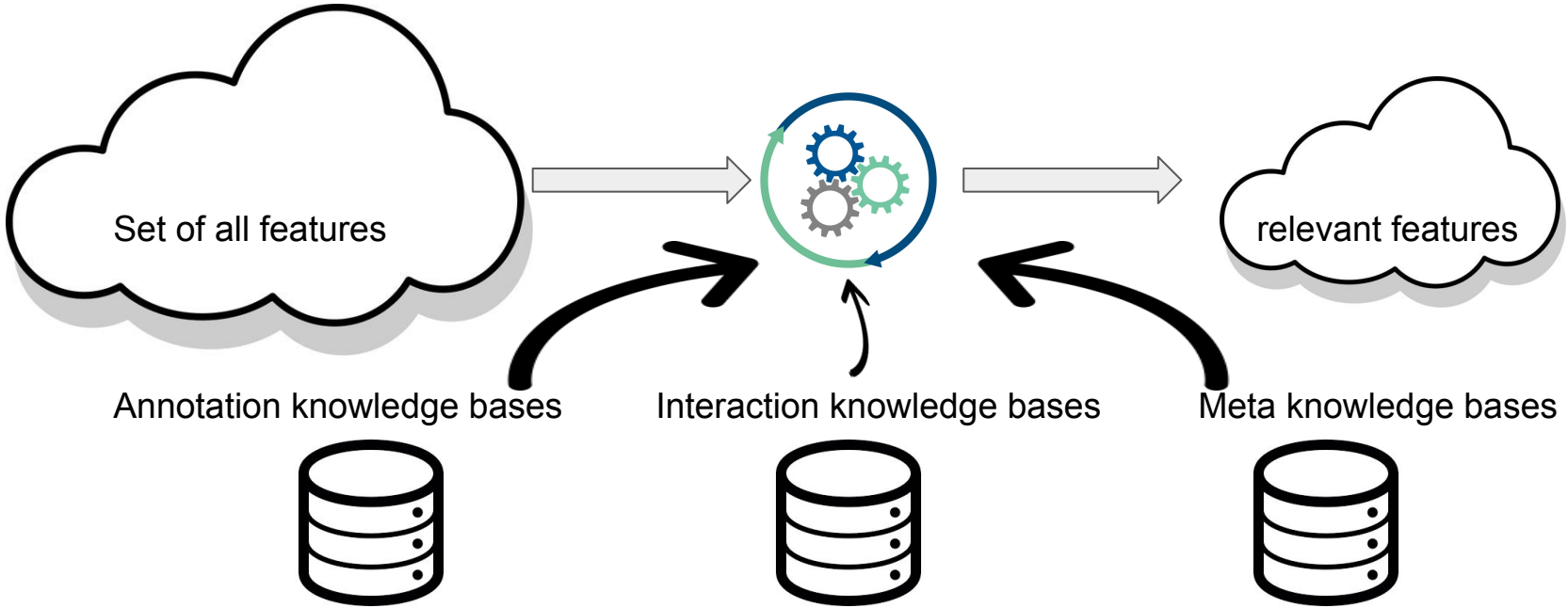
Gene Selection (a.k.a. Feature Selection 4 RNAseq-Data)

- Goals:
  - shorten training times
  - avoid the curse of dimensionality (overfitting)
- Approach:
  - Use statistical models to select interesting “genes” for future analysis
  - Currently use only a bit of existing knowledge of underlying biological processes

Different approaches exist, depending on where and how they integrate external knowledge!

Integrative Gene Selection

Gene Selection using **not only statistical** methods **but also external knowledge bases**!



There is a high demand for

a widely accepted and easily interpretable

Integrative gene selection solution!

What drives us?

- As by today: plenty problems (though list was shortened)
- Knowledge bases grow constantly, reveal new information
- We do not know much regarding biological processes in humans

What drives us? (ctd.)

Hasso Plattner Institut
Digital Engineering · Universität Potsdam

- Especially network approaches for gene selection are very promising!
  ⇒ shown by different studies!

- Compare different integrative approaches
  ⇒ Framework was built already, missing a network-approach

- Goal: participate in offering researchers a great framework for promising research!

How to solve the problems?

Implement Integrative Gene Selection approach

- using network-based gene extraction similar to Gu et al's [2004] approach
- integrate functional categories or functional modules for specific genes via Gene Ontology (GO)'s REST-API

If time is left, I want to integrate biological pathways as features, also from GO, like Quanz et al [2008] have done.

What will be the next steps?

Hasso Plattner Institut
Digital Engineering · Universität Potsdam

1. Implement network-based gene selection solution and integrate KB
   Functional modules and possibly biological Pathways from GO

2. Integrate solution into existing framework IGEA

3. Evaluate solution; compare approach against existing approaches of IGEA

4. Write paper based on findings of achieved work status

Now is the time for Questions and Answers!!!

Hasso Plattner Institut
Digital Engineering · Universität Potsdam

That's it! I am excited to answer your questions or skip this part and start coding right away!

Thank you for your attention!

Image Sources

- DNA Replication: http://www.vcbio.science.ru.nl/en/virtuallessons/cellcycle/trans/
- Chromosome - DNA: https://www.sciencenews.org/article/dna-chromosome-bundles-cell-division-mitosis
- Eukaryote: https://da.wikipedia.org/wiki/Fil:Eukaryote_DNA-en.svg
- mRNA to Protein: https://dumielauxepices.net/wallpaper-3773168
- Gene Ontology: http://www.geneontology.org/page/go-citation-policy
- TCGA: https://www.genome.gov/17516564/the-cancer-genome-atlas/
- WikiPathways: https://en.wikipedia.org/wiki/WikiPathways
- Database: https://www.flaticon.com/free-icon/database_4426
- Cloud: https://all-free-download.com/free-vector/download/nuage-cloud_116075.html
- Process: https://positiveenergy.pro/process/
- Curved Arrow: https://thenounproject.com/term/curved-arrow/69289/

Additional Slides, if needed

Attachment

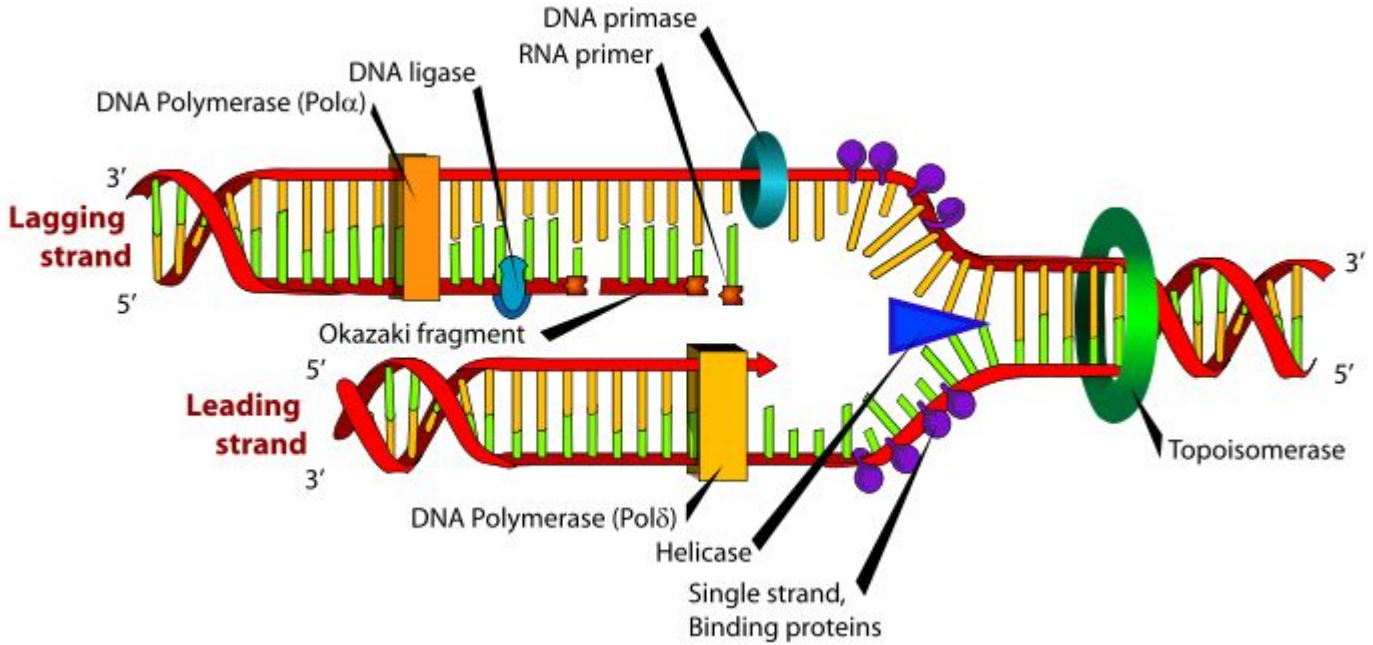
RNA-Seq

RNA-Seq is a recently developed transcriptome profiling technology that utilizes next-generation sequencing platforms (Metzker, 2010; Mardis, 2008). RNA-Seq transcripts are reverse-transcribed into cDNA, and adapters are ligated to each end of the cDNA. Sequencing can be done either unidirectional (single-end sequencing) or bidirectional (paired-end sequencing) and then aligned to a reference genome database or assembled to obtain *de novo* transcripts, proving a genome-wide expression profile (Wang et al., 2009). RNA-Seq offers many advantages over microarray technology. Unlike microarray technology, which depends on already known genes, RNA-Seq is not dependent on existing genome data and can screen novel transcript and analyze transcript structure, including single base-pair resolution and exonic boundaries, which is very valuable while investigating SNPs, thus making it useful for genotyping and linkage analysis (Wang et al., 2009). The advantages of RNA-Seq and its application in studying nervous system and the challenges associated with the technology are summarized in a previous publication (Kadakkuzha and Puthanveettil, 2013).

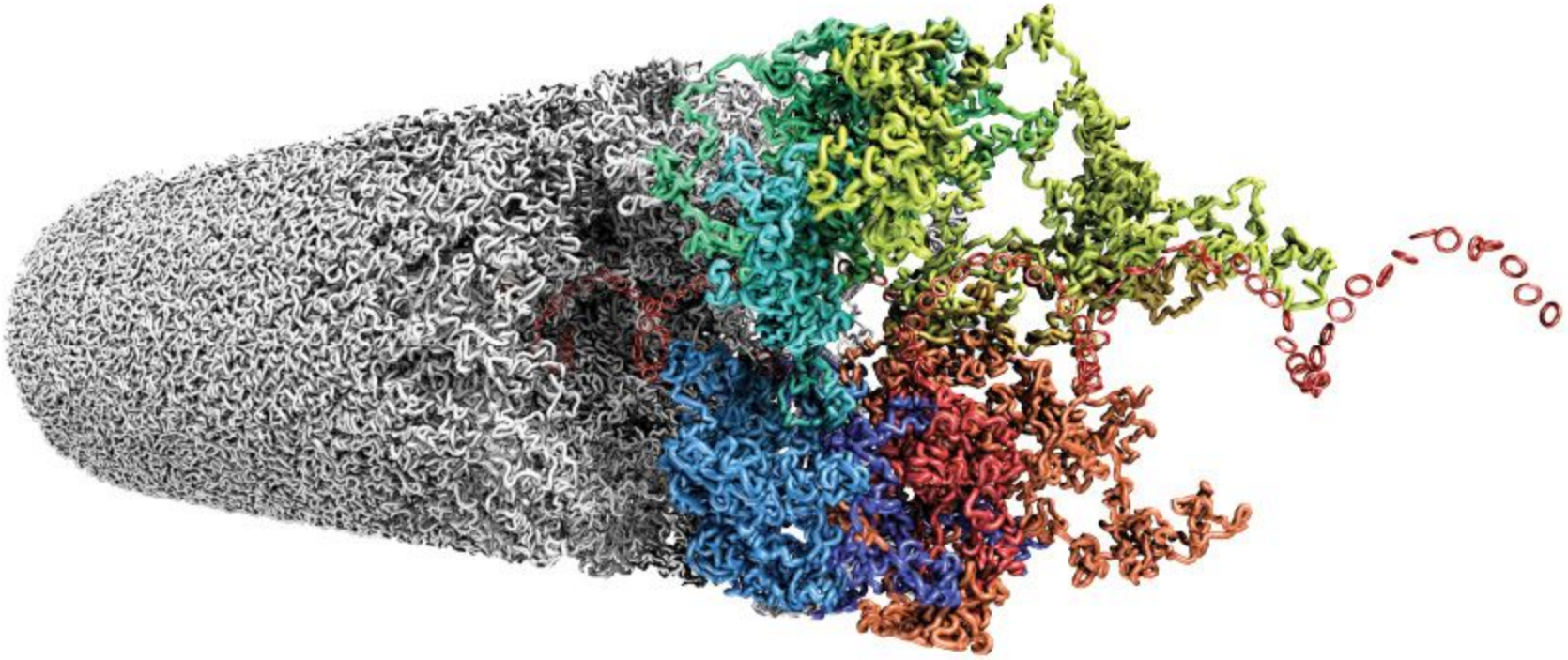
https://www.sciencedirect.com/topics/neuroscience/rna-seq

Problems only statistic approaches are facing (ctd.)

- Deficiencies when it comes to robustness and stability aross data sets and approaches, due to the statistical issue of working in a high-dimensional space with only a few samples; **Low Robustness and Stability**
- The more complex an approach is, the less trustworthy it becomes, due to users considering them a black box, especially when machine learning techniques are used; **Lack of Transparency**
- Genes sharing similar functions, participating in the same pathway, often share similar expression patterns, ⇒ dependencies in expression profiles; **Redundancy**

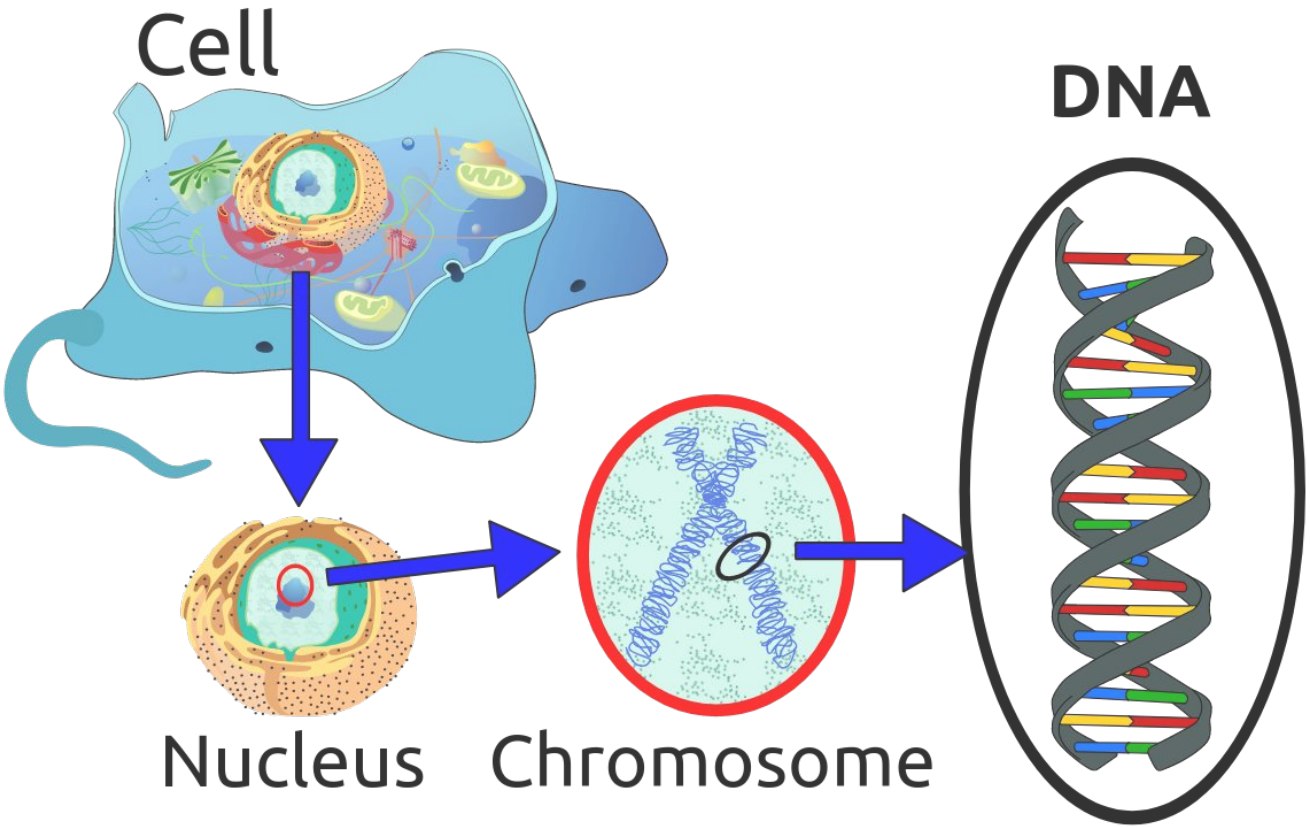


Source: http://www.vcbio.science.ru.nl/en/virtuallessons/cellcycle/trans/

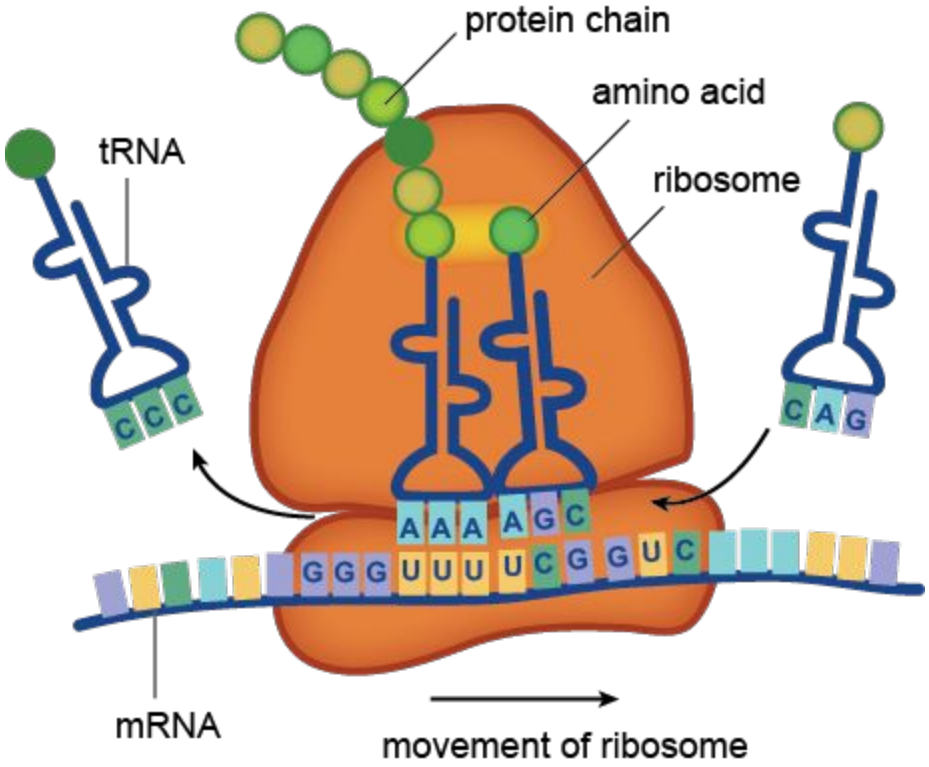


Source:

https://www.sciencenews.org/article/dna-chromosome-bundles-cell-division-mitosis



Source: https://da.wikipedia.org/wiki/Fil:Eukaryote_DNA-en.svg



Source: https://dumielauxepices.net/wallpaper-3773168

Challenges for integrative gene selection

Genes can be involved in multiple pathways; **Pathway Overlaps**

- Gap between static interaction networks and dynamic cell processes; **Conceptional Gap**

- Knowledge bases heavily vary in quality of information, ranking scores, review status of information, evidence, disease specificity and many more aspects; **Knowledge Base Quality**

Challenges for integrative gene selection

Hasso Plattner Institut
Digital Engineering · Universität Potsdam

- It is a fine line between RNAseq data results and integrated external knowledge; the most advanced approach for integrative gene selection has no added value if it cannot be appliced in practice due to confusing setup-processes, specific use-cases and so on; updates of knowledge bases sometimes require source-code maintenance / refactorings of integrative gene selection approaches; **Gene Selection Process**

- Current knowledge on biological processes and diseases is not exhaustive, thus knowledge bases are updated frequently, resulting in reproducibility issues as time goes by; **Evaluation**

Motivation for this specific topic / approach

- Untersch. Integrative Ansätze miteinander vergleichen, Network-Based fehlt
- Network-Based Integrative Gene Selection achieved rece
- Anbindung weiterer Knowledge-Bases an das bestehende Framework, um ein generalisierbares Framework für verschiedenste Ziele zu erstellen und Wissenschaftlern an die Hand zu geben, um Ihre eigenen Ansätze mittels dieses Frameworks bewerten zu lassen (Vielleicht einen eigenen Ansatz, der div. Ansätze kombiniert, entwickeln?)

ToDo’s:

- Shorten up text @ bulletpoints, only use main keywords
- When a new variant of slides is ready, send to Cindy

Das hier wird wahrscheinlich zu viel...

Genome sizes are typically given as gametic nuclear DNA contents (‘C-values’) either in units of mass (picograms, where 1 pg = 10^{-12} g) or in number of base pairs (in eukaryotes, most often in megabases, where 1 Mb = 10^6 bases). These are directly interconvertible as 1 pg = 978 Mb (or 1 Mb = 1.022×10^{-3} pg) (7).

Genome Size of humans: 3.2 Gbp

Genes: approx. 20k-25k

Mitosis & Meiosis - Cell Division (Eukaryotes vs. Gamete)