CMSC 435
Intro to Data Science

Lecture set 7. Case Study

Instructor: Lukasz Kurgan

Introduction

Assignments 2 and 3 and project concern analysis of a real-life dataset

- practical application of data science
- we will use the same data but make it gradually more challenging as the semester progresses

Goals for today’s lecture:

- describe background behind these data
- motivate and explain the analysis that will be done in the assignments and project

Overview

Case study involves computational analysis of data about proteins

- application of data science in modern molecular biology

Outline

- short introduction to proteins and protein data
- motivation to use data science
- practical hints for the assignment 3 and project

Introduction to Proteins

From the Greek *protas* meaning *of primary importance*

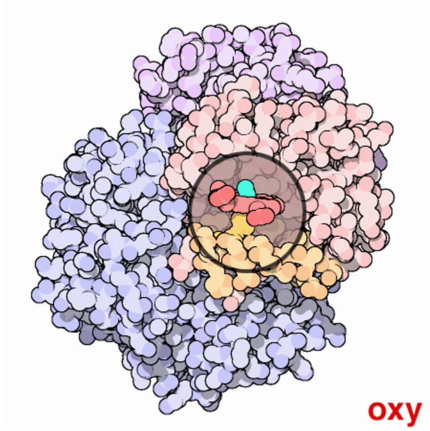
Organic molecules that are composed of a sequence of amino acids

Arguably the most actively-studied biological macromolecules

- other biological macromolecules include DNA, various types of RNA, polysaccharides and lipids

Human hemoglobin (subunit alpha)

MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF
DLSHGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRV
DPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR



Amino Acids	**Abbr.**	**Hydro-phobic**	**Charge**	**Size**		**Occurrence (%)**
				Small	**Tiny**	
Alanine	Ala, **A**	X		X	X	7.8
Cysteine	Cys, **C**	X		X		1.9
Aspartate	Asp, **D**		negative	X		5.3
Glutamate	Glu, **E**		negative			6.3
Phenylalanine	Phe, **F**	X				3.9
Glycine	Gly, **G**	X		X	X	7.2
Histidine	His, **H**		positive			2.3
Isoleucine	Ile, **I**	X				5.3
Lysine	Lys, **K**		positive			5.9
Leucine	Leu, **L**	X				9.1
Methionine	Met, **M**	X				2.3
Asparagine	Asn, **N**			X		4.3
Proline	Pro, **P**	X		X		5.2
Glutamine	Gln, **Q**					4.2
Arginine	Arg, **R**		positive			5.1
Serine	Ser, **S**			X	X	6.8
Threonine	Thr, **T**	X		X		5.9
Valine	Val, **V**	X		X		6.6
Tryptophan	Trp, **W**	X				1.4
Tyrosine	Tyr, **Y**	X				3.2

Introduction to Proteins

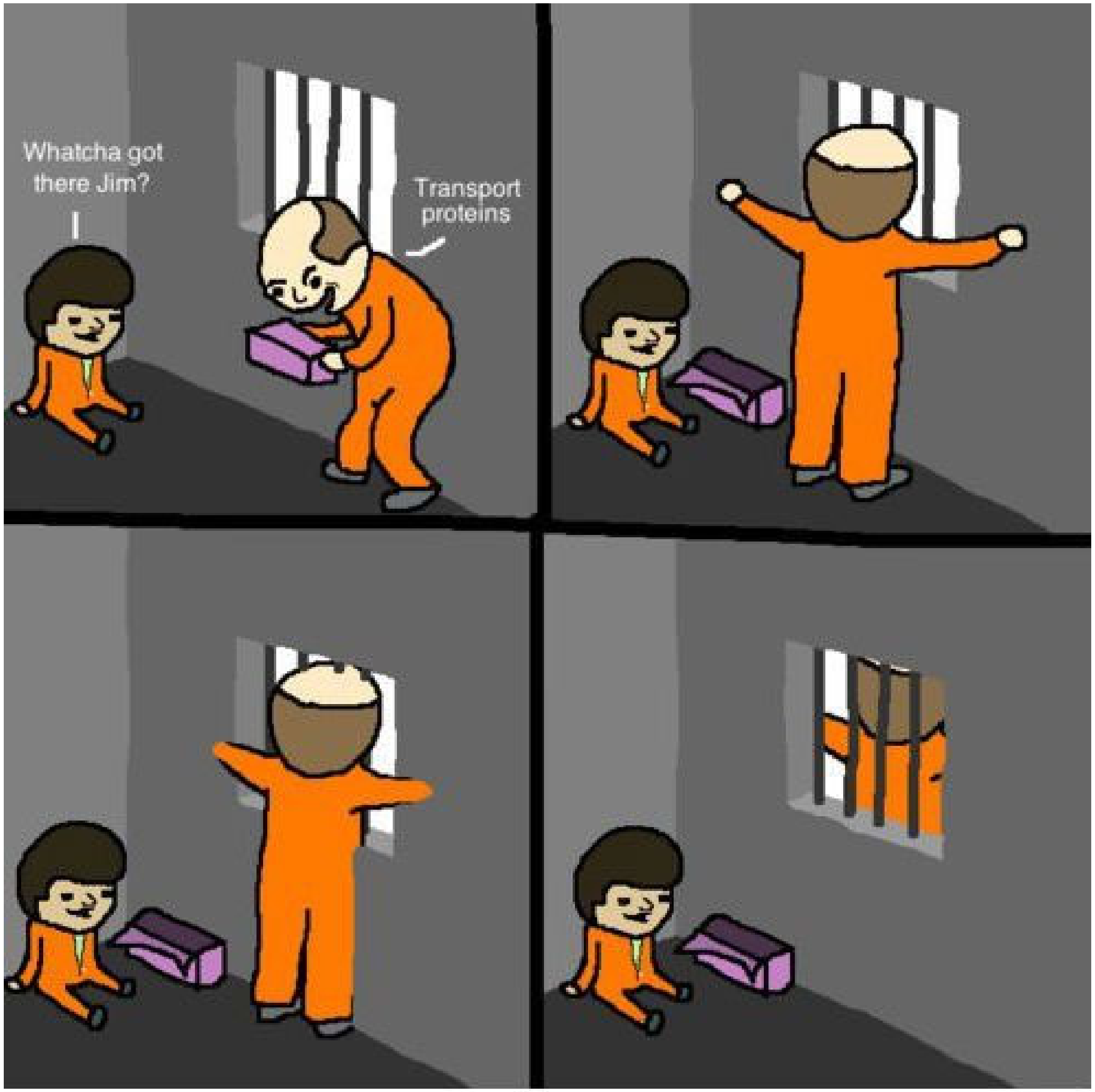
Essential to virtually all aspects of life

- catalyze chemical reactions (enzymes), form cytoskeleton (tubulin), provide signaling and transporting functions (hemoglobin), implement immune responses (antibodies), pack DNA in cells (histones), regulate cell processes (hormones) and the list goes on...
- 70,000 different protein types in human
  - each protein adopts a different shape in spite of being composed of the same 20 amino acids
  - each protein is replicated billions of times across cells

Proteins “do” everything

Introduction to Proteins

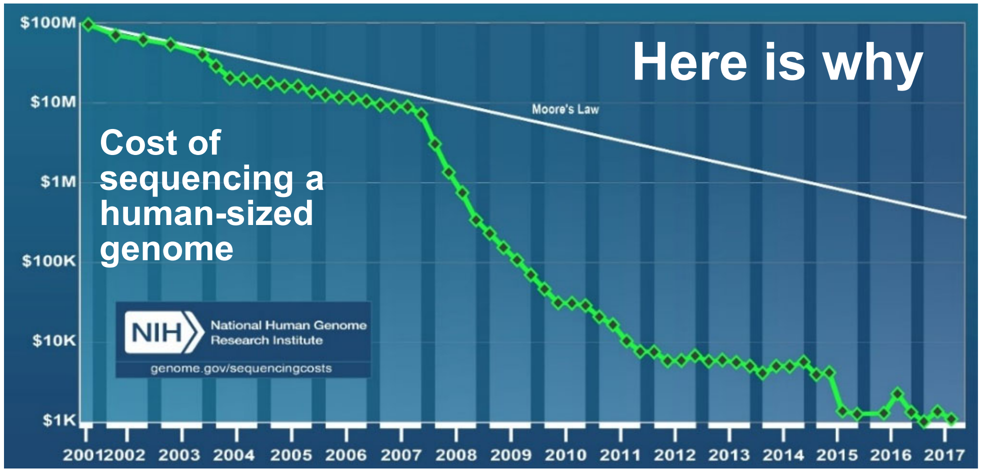
Transport proteins move many different molecules across cell walls (membranes)

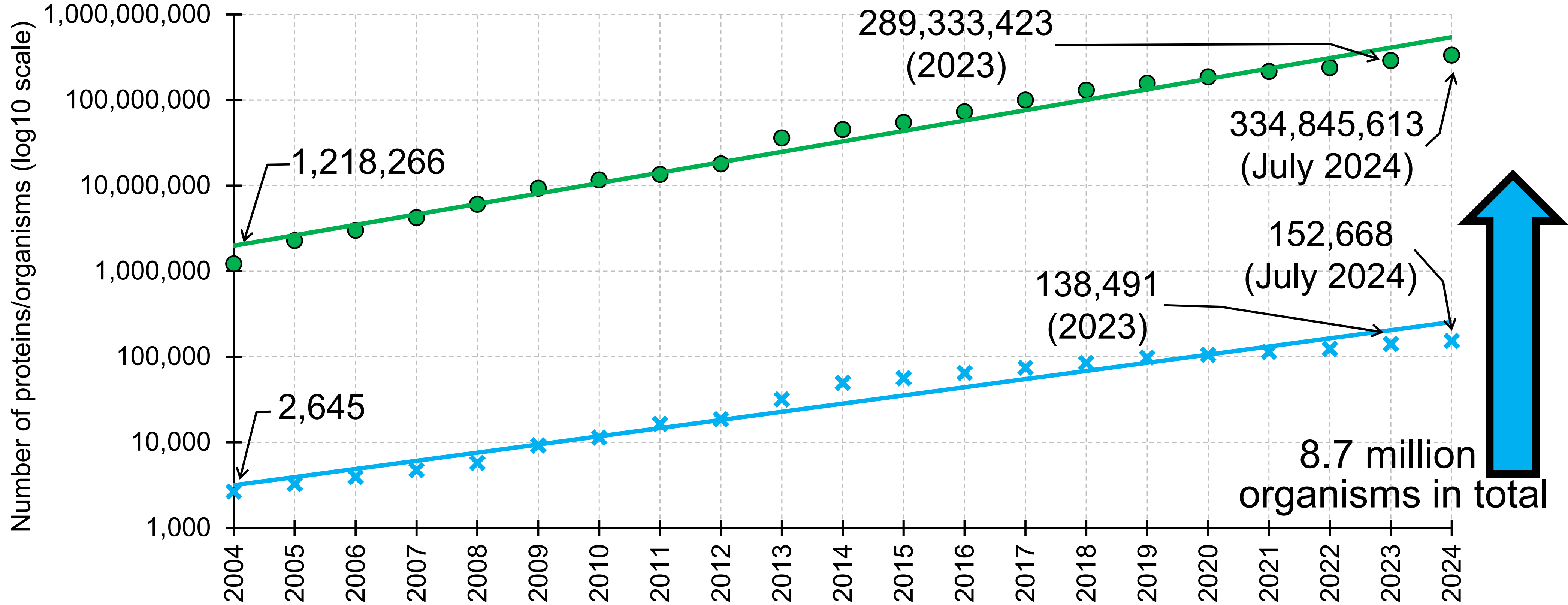


Taken from http://weknowmemes.com/2013/04/transport-proteins-comic/

© Lukasz Kurgan, 2024

BIG data





Source: RefSeq database at https://www.ncbi.nlm.nih.gov/refseq/

Exponential growth in the number of known protein sequences is a big challenge since we don't know what over 99% of them do

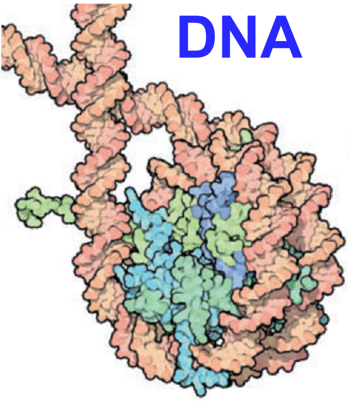
© Lukasz Kurgan, 2024

Introduction to Proteins

Functions of proteins result from interactions with other molecules

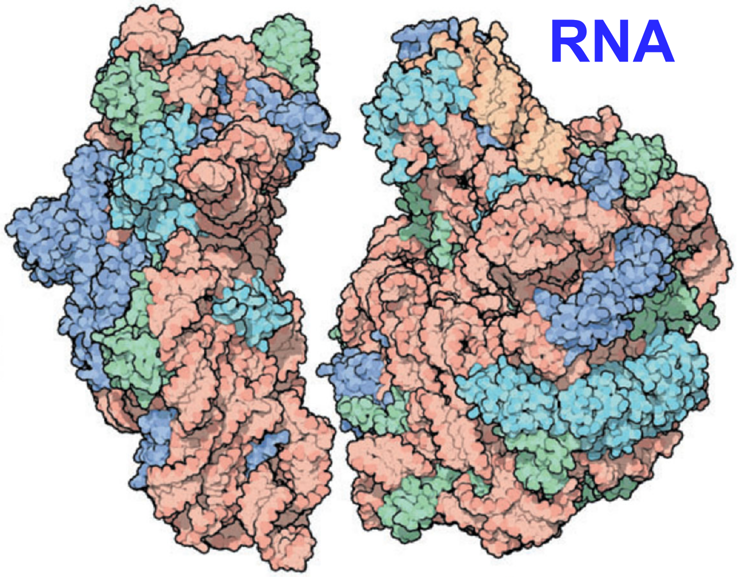
Protein sequence determines its function, i.e., function can be predicted from sequence

DNA



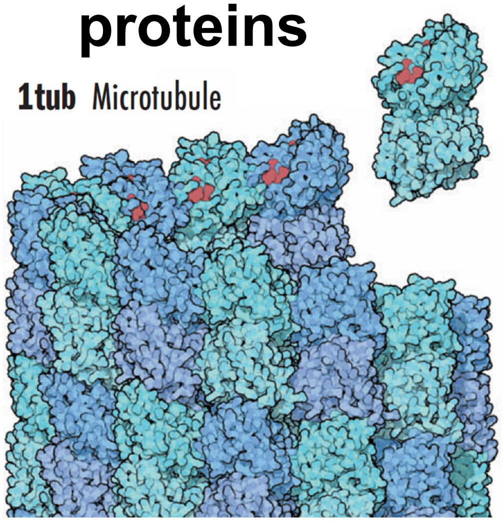
1aoi Nucleosome

RNA



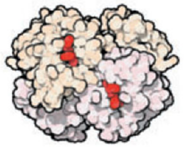
1fjf+1jj2 Ribosome

other proteins



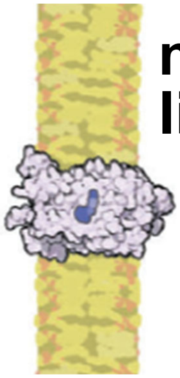
1tub Microtubule

small molecules



4hhb Hemoglobin

membranes lipids



1f88 Rhodopsin

Introduction to Proteins

We focus on proteins that interact with RNA and DNA

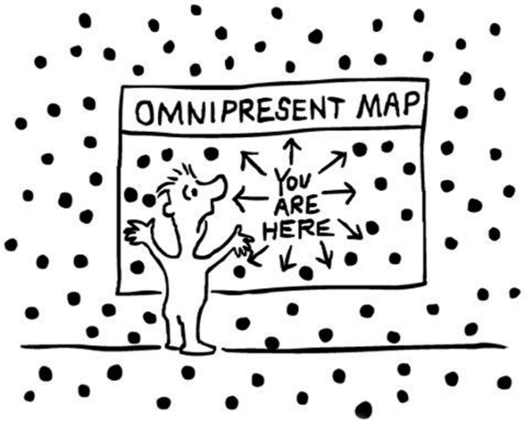
- 3% of proteins in eukaryotic organisms interact with DNA
- among the currently sequenced 80.4 million eukaryotic proteins we should find

  3% of 80.4 million = 2,409,000 that interact with DNA
- so far we found only about

  120,000 (4.9%) of them

- finding the remaining 95% at the current pace would take decades, while in the meantime the number of protein sequences will continue to grow exponentially

Introduction to Proteins



To summarize

- proteins are important and omnipresent
- we know 335 million of their sequences but we do not know what vast majority of them do

Accurate prediction of interaction with DNA or RNA from a sequence would be helpful to decipher function of that protein sequence

- this must be performed computationally given the large and continually growing protein data

Can we use data science to do that?

Yes, we can

Use historical data to build predictive model, test it, and if the model offers good predictive performance then use it to predict new data

historical data: proteins known to interact with RNA, DNA, and known not to interact with the nucleic acids

predictive model: takes protein sequence as the input and outputs whether or not it can interact with DNA and/or RNA

predictive performance: use the model to predict a set of proteins for which we know the outcomes and compare the predicted with the true/known outcomes

Assignment 2

You have performed imputation of missing values for a dataset of proteins

- **each object (row in the dataset) is a protein sequence**
- **sequences are represented by a small set of provided features**

- **in the mean imputation approach we used average protein feature to fill-in missing values**
- **in the hot deck imputation approach we used the most similar other protein to fill-in missing values**

Assignment 3

Predict whether a given protein sequence would / would not interact with nucleic acids (either DNA or RNA)

- sequences are represented by a small set of provided features
- proteins are labelled as Yes (interact with DNA or RNA) and No (do not interact)
- evaluate and compare predictive performance for a pre-determined set of models using the dataset of labelled proteins
- compare models built using different algorithms and evaluate using different types of tests

Self-contained individual work

- the next class will introduce RapidMiner (Altair AI Studio) software for building and assessment of predictive models

Groups of 6 or 7 students

Project

Predict whether a given protein sequence would interact with DNA, RNA, both, or neither

- use the (provided) dataset of proteins annotated as "DNA", "RNA", "DNA and RNA", and "neither" to build "open-ended" predictive models and evaluate their performance in more detail
- proteins are represented by features that you will have to generate from the protein sequences
- compare models built using different designs (algorithms, features, parameters) on the provided dataset to identify the best solution
- evaluate the selected solution using a new (blind) dataset
- make a group presentation that describes how your model works and discusses your results

Groups of 6 or 7 students

Project

Project steps (using the 6-step KD process)

3. Preparation of the data

design and compute features from the input sequence (create rectangular dataset); perform data cleaning, feature/example selection and/or transformation

4. Data Mining

select and setup specific architecture of your model; consider alternatives

calculate specific criteria to measure predictive performance

perform specific types of tests

5. Evaluation of the discovered knowledge

understand and discuss your results

compare your results to existing method(s)

6. Using the discovered knowledge

write the report

predict proteins on the blind test dataset

make the presentation

Groups of 6 or 7 students

Project

Hints

Resources to encode sequences into features

- Pfeature (code in python + webserver)
  https://github.com/raghavagps/Pfeature
- iLearnPlus (code in python + webserver)
  https://github.com/Superzchen/iLearnPlus
- MathFeature (code in python + webserver)
  https://github.com/Bonidia/MathFeature
- AAindex database
  https://www.genome.jp/aaindex/
  use AAindex1 to encode amino acids using their physicochemical properties and next aggregate (average) these properties per sequence

Ensure that you will be able to efficiently duplicate this process for the test dataset when it becomes available (even if online resources would go down…)

Groups of 6 or 7 students

Project

Hints

- using too many features may hurt the predictive quality and lead to overfitting (over-describing) the dataset
  - use feature selection if you end up generating too many features
- rationally/empirically choose the best architecture for your model
  - parameterize your selected algorithm(s); use multiple models together
- consider different ways to perform predictions
  - e.g., one predictor for 4 outcomes vs. 4 predictors, each for 1 outcome

Project dates

- release on October 15 (after midterm exam)
- progress reports on November 7
- final reports due on November 22
- presentations sessions on December 3 and 5

More in the project document on October 15