CMSC 635
Knowledge Discovery and Data Mining

Lecture set 14. Project description

Instructor: Lukasz Kurgan

Overview

Case study involves computational analysis of protein data

Outline

- short (and painless) introduction to proteins
- why data mining
- project description
- hints for the project

Introduction to Proteins

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

Organic molecules formed by a sequence of amino acids

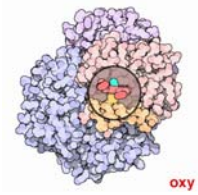
Arguably the most actively-studied biological macromolecules

- **other biomacromolecules include DNA, various RNAs, polysaccharides and lipids**

3

Human hemoglobin subunit alpha

MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF
DLSHGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRV
DPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR



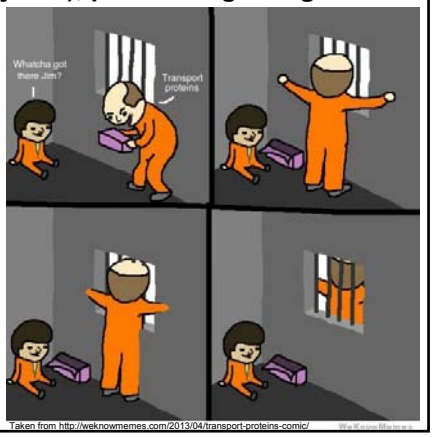
Amino Acids	Abbr.	Hydro-phobic	Charge	Size		Occurrence (%)
				Small	Tiny	
Alanine	Ala, **A**	X		X	X	7.8
Cysteine	Cys, **C**	X		X		1.9
Aspartate	Asp, **D**		negative	X		5.3
Glutamate	Glu, **E**		negative			6.3
Phenylalanine	Phe, **F**	X				3.9
Glycine	Gly, **G**	X		X	X	7.2
Histidine	His, **H**		positive			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

4

Introduction to Proteins

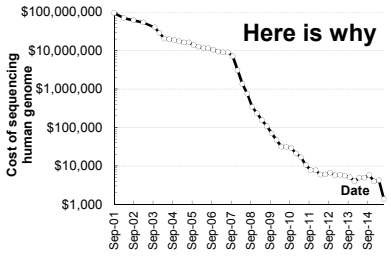
Essential to the structure and function of all species

- **catalyze chemical reactions (enzymes), perform signaling and transporting functions (hemoglobin; aquaporin), implement immune responses (antibodies), regulate cell processes (hormones; transcription factors), ...**
- **about 70,000 different human proteins**
- **aside from the fat, human body has about 20% of proteins by weight**



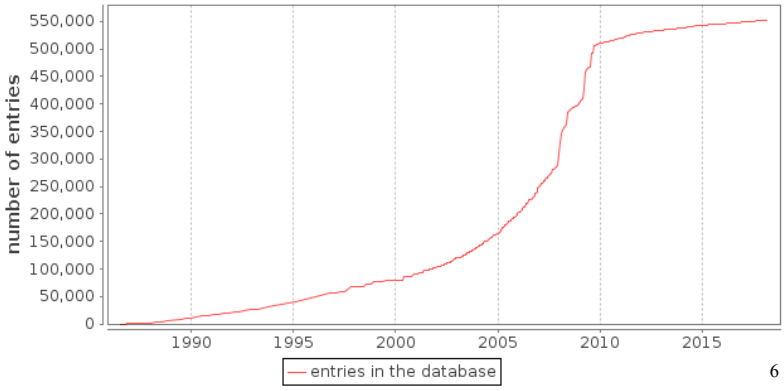
Taken from http://weknowmemes.com/2013/04/transport-proteins-comic/

Introduction to Proteins



So, what is the problem?

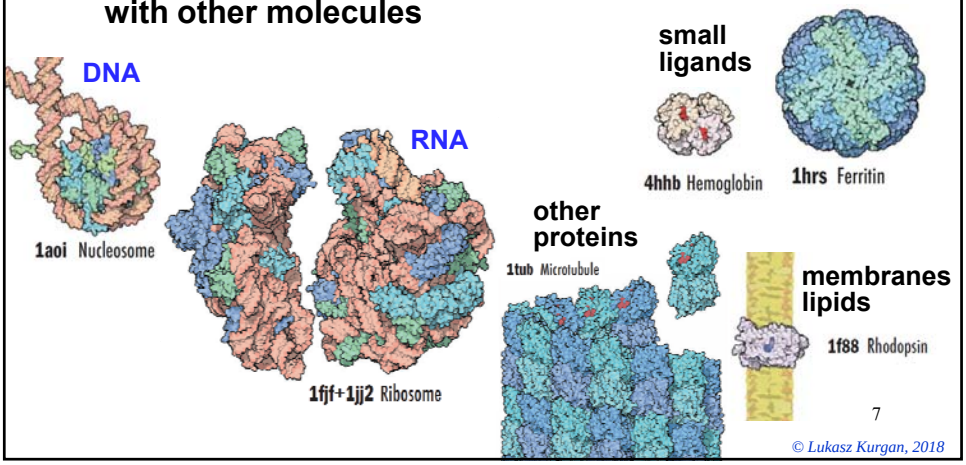
Number of entries in UniProtKB/Swiss-Prot over time



6

Introduction to Proteins

Functions of proteins result from interactions with other molecules



7

Introduction to Proteins

We focus on proteins that interact with RNA and DNA

- **the abundance of DNA-binding proteins was recently estimated to be on average at 3% of proteins in eukaryotic organisms and 5% in the animals**
- **we sequenced 27 million eukaryotic proteins** (source: UniProt as of March 2, 2018) **and assuming conservative estimates we expect to know**

3% of 27 million =

810,000 DNA-binding eukaryotic proteins

- **so far we know only about**

45,000 (5.5%) DNA-binding eukaryotic proteins

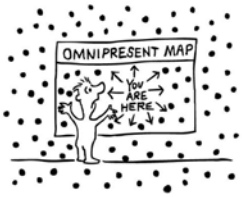
when including both experimental data and predictions

8

Introduction to Proteins

To summarize

- **proteins are important and omnipresent**
- **we know 100+ million of their sequences but not what most of them do**



Ability to predict whether a given sequence would interact with DNA or RNA would be very helpful to decipher function of that protein

- **this must be performed computationally given the large and continually growing scale of the problem**

Can we use data science to do that?

9

Yes

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

data: proteins annotated as interacting with RNA, DNA, and not interacting with 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

10

Project

Goal is to accurately predict whether a given protein sequence interacts with DNA, RNA, both or neither.

Use empirical results on a training dataset to select and optimize the best design

- **compare different designs (algorithms, features, parameters) on the training dataset**

Use the best design to perform predictions on a "blind" test dataset

- **instructor will empirically assess these results**

Project

Project steps

3. Preparation of the data

design and compute features from the input sequence (create rectangular dataset); perform feature 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

Project

Hints

Resources to encode sequences into features

- PROFEAT server
  http://bidd2.nus.edu.sg/cgi-bin/profeat2016/protein/profnew.cgi
- ProtrWeb
  http://protrweb.scbdd.com/
- ProtDCal
  http://bioinf.sce.carleton.ca/ProtDCal/
- AAindex database
  http://www.genome.jp/aaindex/
  - use AAindex1 to encode amino acids using their (selected) physicochemical properties and next aggregate these properties per sequence

Remember to ensure that you will be able to efficiently duplicate this process for the test dataset when it becomes available

Project

Hints

- using too many features may hurt the predictive quality and lead to overfitting (over-describing) the dataset
  - use feature selection if needed (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 4-class predictor vs. 4 binary predictors
- understand and explain what you are doing

More in the project document

You will have five weeks to complete the project

- start early, use the time wisely, and **GOOD LUCK**