Laboratory for Bioinformatics and Computational Chemistry
Institute of Nuclear Sciences VINCA

Prediction of protein functions and protein-protein interactions using machine learning

Vladimir Perović

MLA@MATF

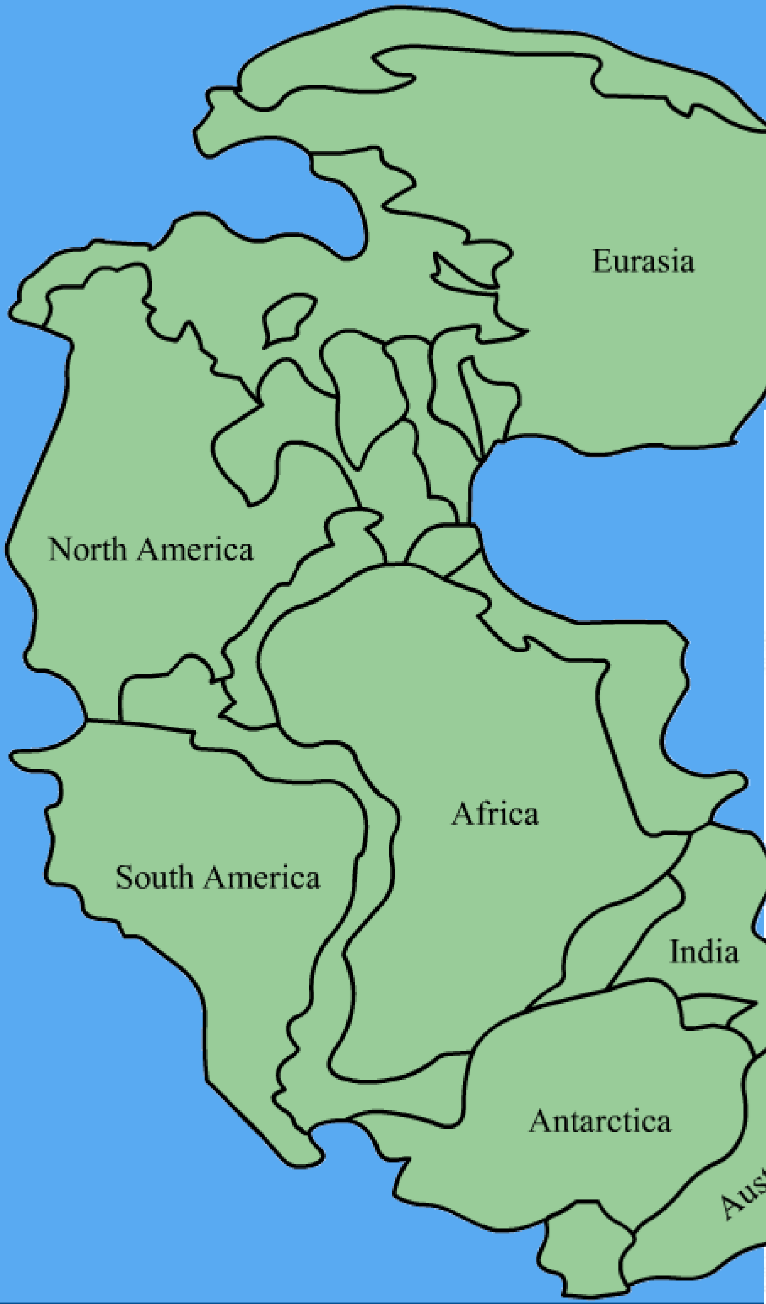
Laboratory for Bioinformatics and Computational Chemistry
Institute of Nuclear Sciences VINCA

Prediction of protein functions and protein-protein interactions using machine learning

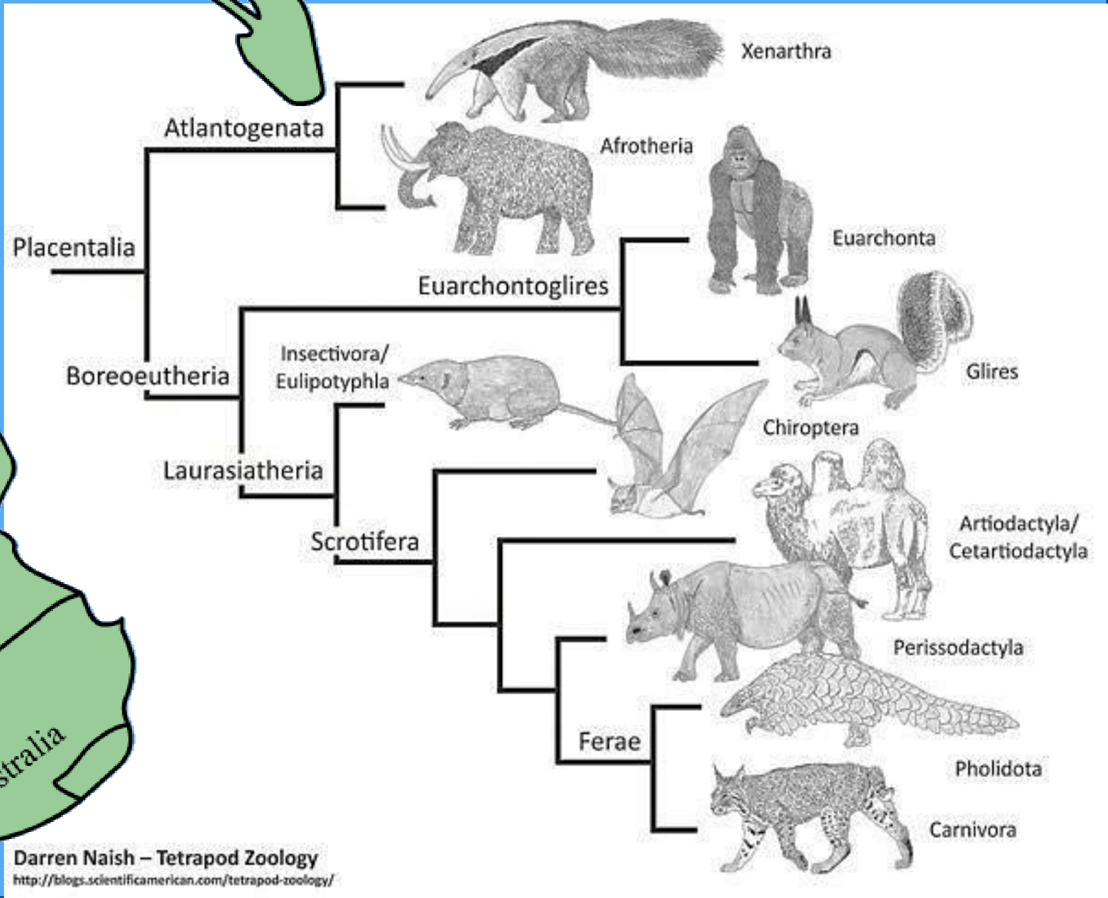
Vladimir Perović

MLA@MATF

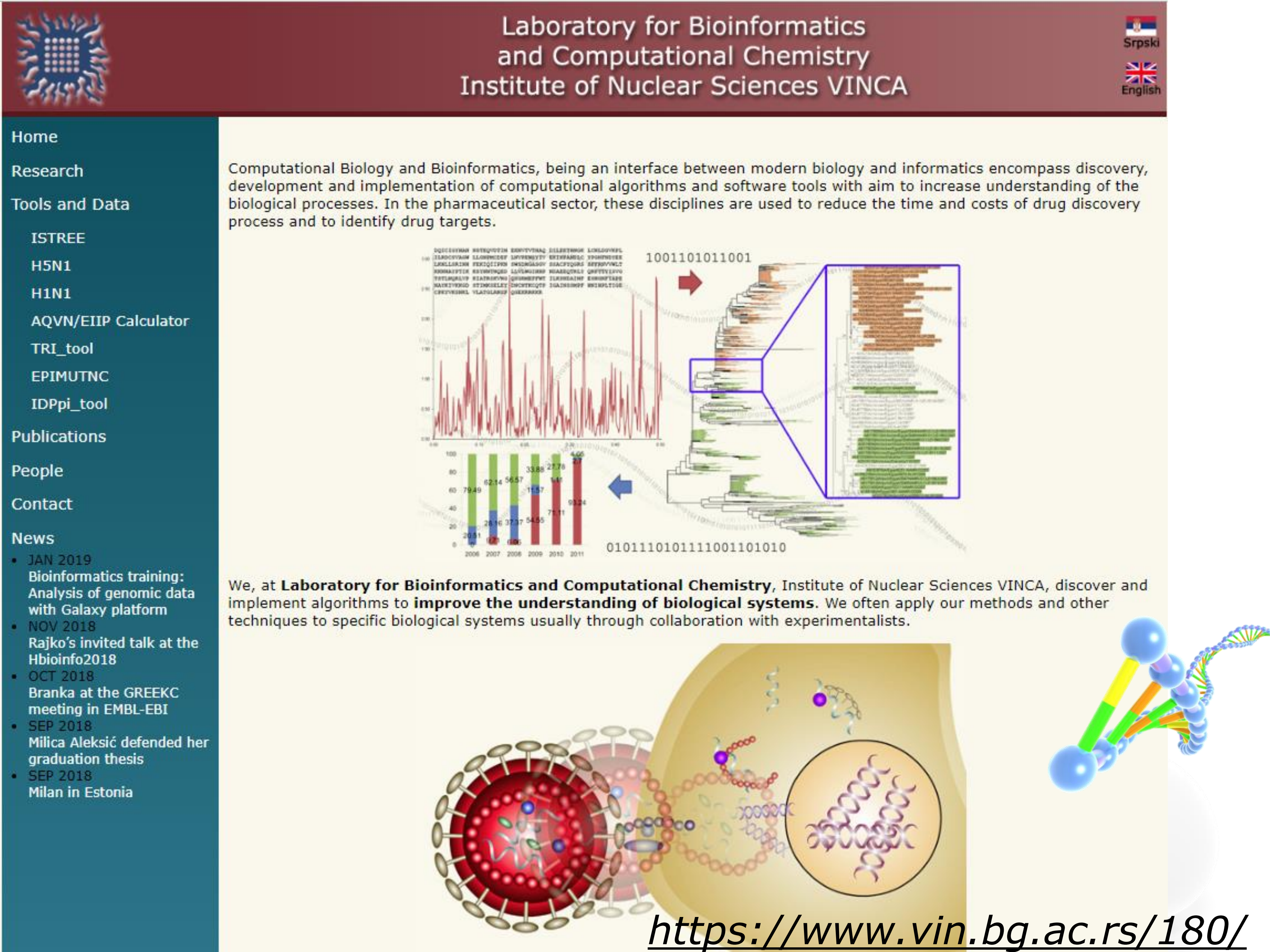
PANGEA supercontinent



Molecular phylogenetic placental tree



Darren Naish – Tetrapod Zoology
http://blogs.scientificamerican.com/tetrapod-zoology/



Laboratory for Bioinformatics
and Computational Chemistry
Institute of Nuclear Sciences VINCA

News

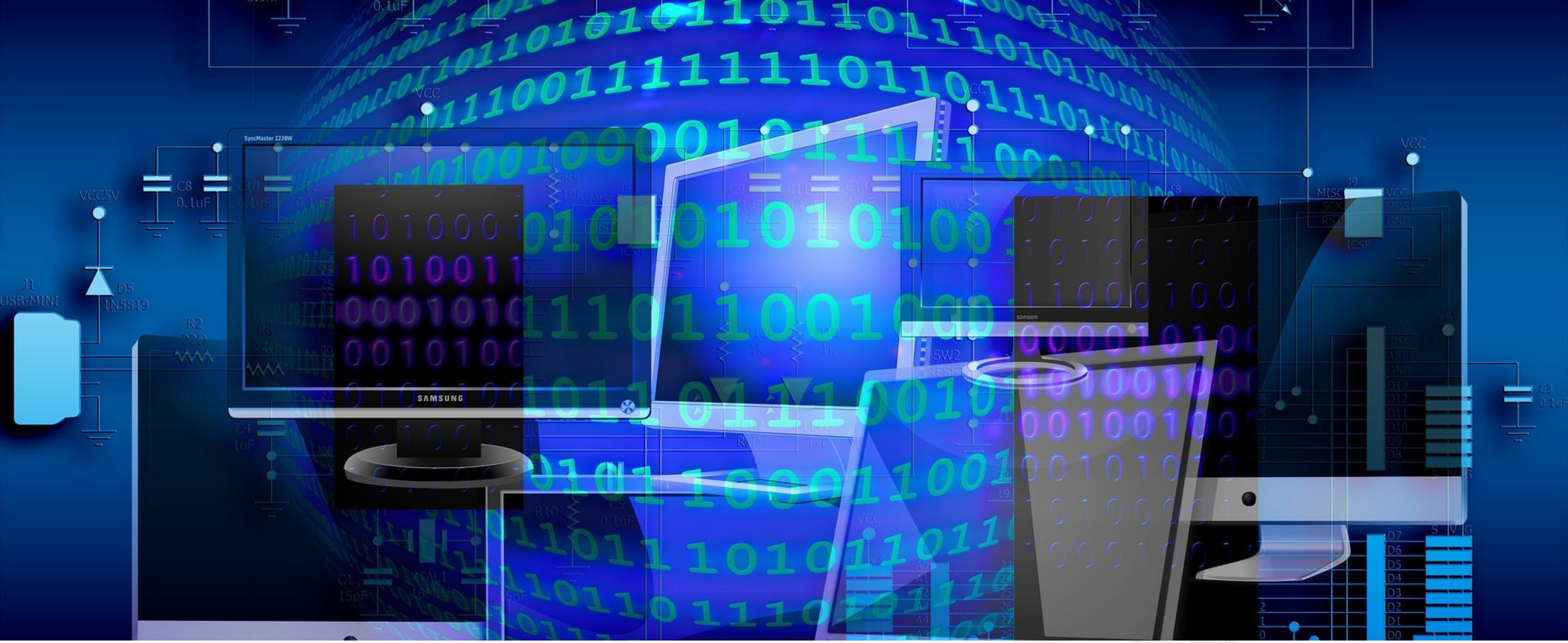
- JAN 2019
  Bioinformatics training: Analysis of genomic data with Galaxy platform
- NOV 2018
  Rajko's invited talk at the Hbioinfo2018
- OCT 2018
  Branka at the GREEKC meeting in EMBL-EBI
- SEP 2018
  Milica Aleksić defended her graduation thesis
- SEP 2018
  Milan in Estonia

Computational Biology and Bioinformatics, being an interface between modern biology and informatics encompass discovery, development and implementation of computational algorithms and software tools with aim to increase understanding of the biological processes. In the pharmaceutical sector, these disciplines are used to reduce the time and costs of drug discovery process and to identify drug targets.

We, at **Laboratory for Bioinformatics and Computational Chemistry**, Institute of Nuclear Sciences VINCA, discover and implement algorithms to **improve the understanding of biological systems**. We often apply our methods and other techniques to specific biological systems usually through collaboration with experimentalists.

https://www.vin.bg.ac.rs/180/

We're going from an information age to a knowledge age



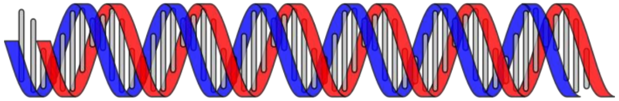
Prediction of protein functions and interactions using machine learning algorithms

Outline

- **What are proteins?**
  - Structure, function, sequence and interactions
- **Protein-protein interaction prediction problem**
  - Important and challenging?
  - PPI algorithm evaluation
  - Our proteome-wide approach
  - Our class-specific approaches (IDPs, TRFs)
- **Protein function prediction problem**
  - Ontological annotation of proteins
  - Gene ontologies and CAFA challenge
  - Our proteome-wide HPO prediction

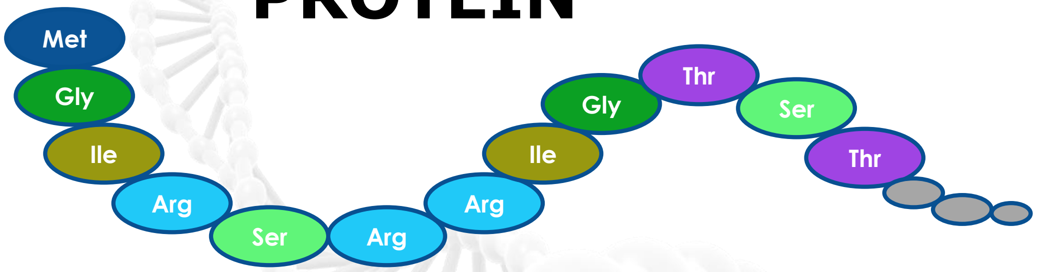
What are Proteins?

DNA



Translation

PROTEIN



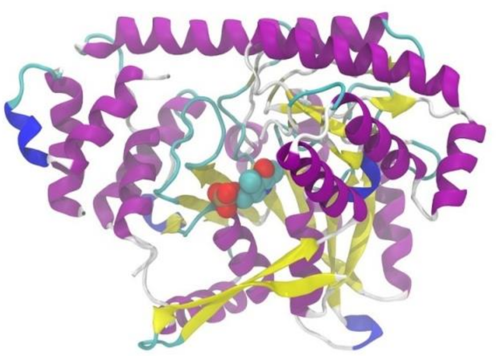
Protein sequence

```
>INS_HUMAN Insulin OS=Homo sapiens
MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSH
LVEALYLVCGERGFFYTPKTRREAEDLQVGQVEL
GGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSL
YQLENYCN
```

Folds

Function

Carries



3D structure

Protein Sequence

Sequence = String

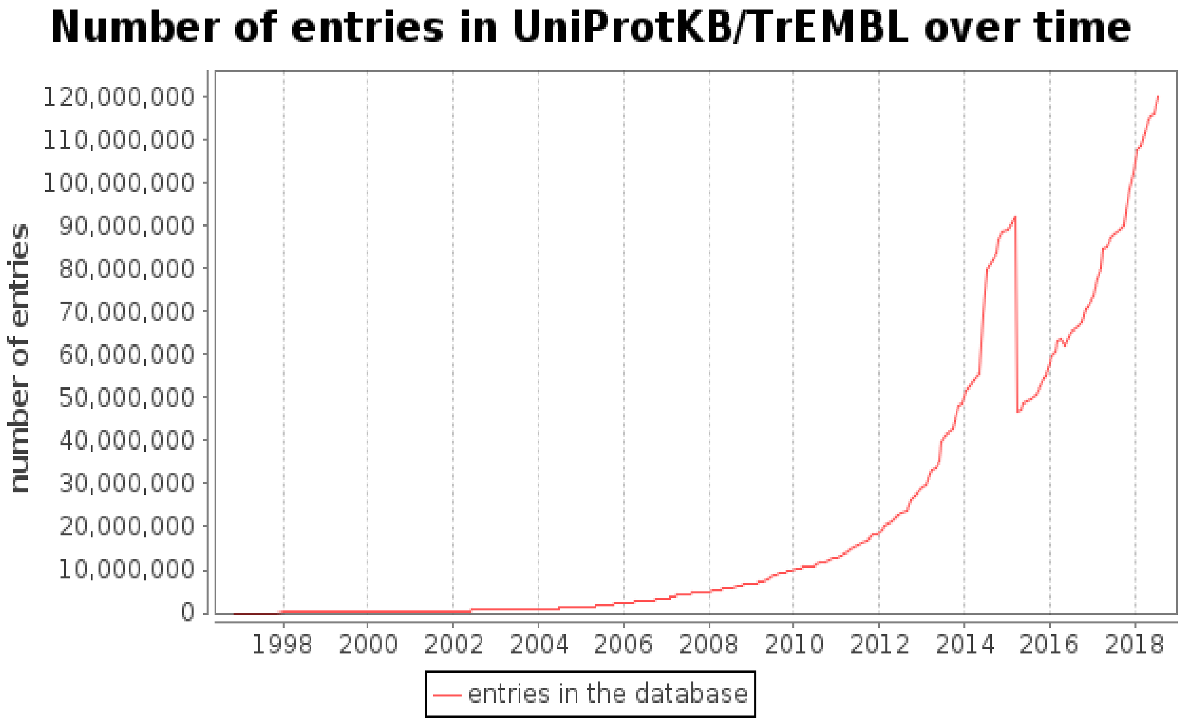
- 20 amino acids; 20 symbols

{A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V, W,Y}

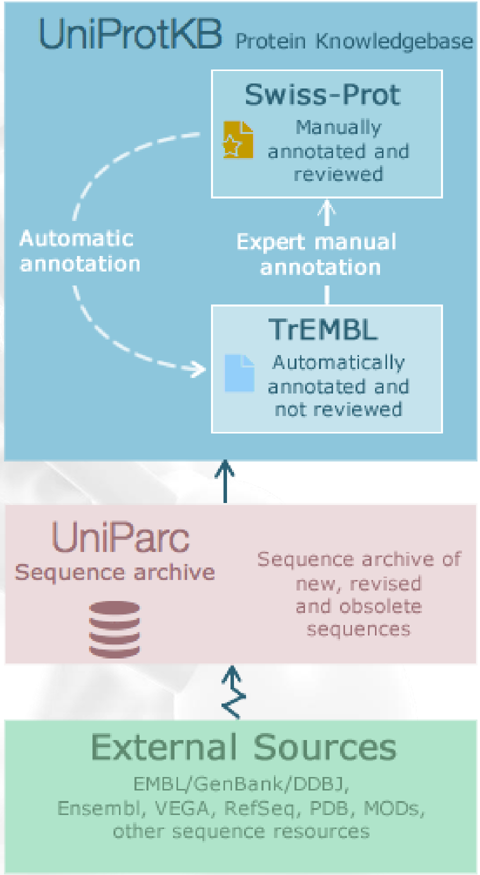
```
MATAERRALGIGFQWLSLATLVLICAGQGGRREDGGPACYGGFDLY
FILDKSGSVLHHWNEIYYFVEQLAHKFISPQLRMSFIVFSTRGTTL
MKLTEDREQIRQGLEELQKVLPGGDTYMHEGFERASEQIYYENRQG
YRTASVIIALTDGELHEDLFFYSEREANRSRDLGAIVYCVGVKDFN
ETQLARIADSKDHVFPVNDGFQALQGIIHSILKKSCIEILAAEPST
ICAGESFQVVVRGNGFRHARNVDRVLCSFKINDSVTLNEKPFSVED
TYLLCPAPILKEVGMKAALQVSMNDGLSFISSSVIITTTHCSDGSI
LAIALLILFLLLALALLWWFWPLCCTVIIKEVPPPPAE
```

UniProt Universal Protein resource, a central repository of protein data

120,243,849 sequence entries



Number of entries in UniProtKB/TrEMBL over time

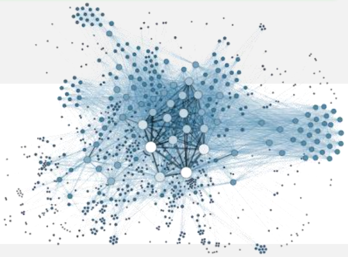


UniProtKB Protein Knowledgebase
Swiss-Prot
Manually annotated and reviewed
Automatic annotation
Expert manual annotation
TrEMBL
Automatically annotated and not reviewed
UniParc
Sequence archive
Sequence archive of new, revised and obsolete sequences
External Sources
EMBL/GenBank/DDBJ, Ensembl, VEGA, RefSeq, PDB, MODs, other sequence resources

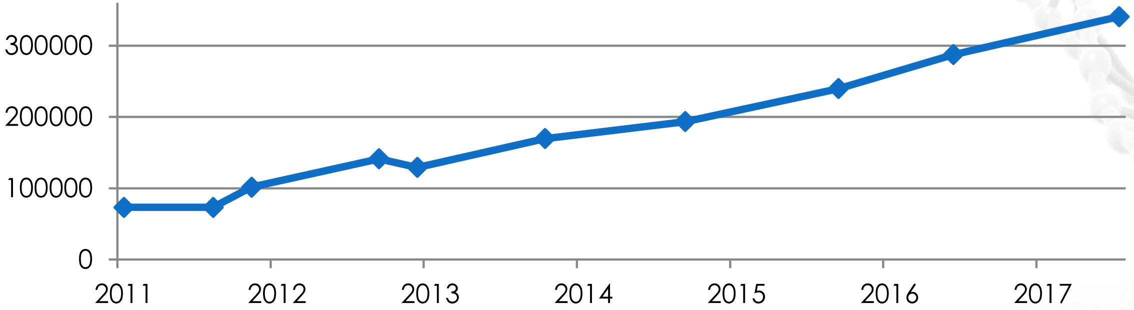
Molecular Interactions

Protein protein interaction (PPI) network = Graph

- Node = Protein
- Link = Bind or carry out same function

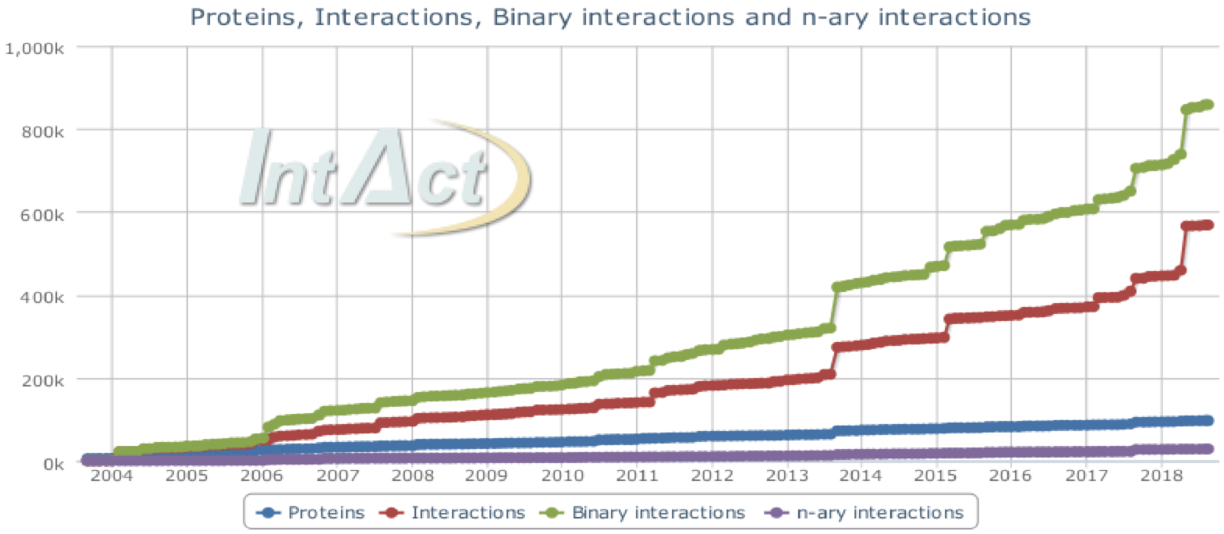


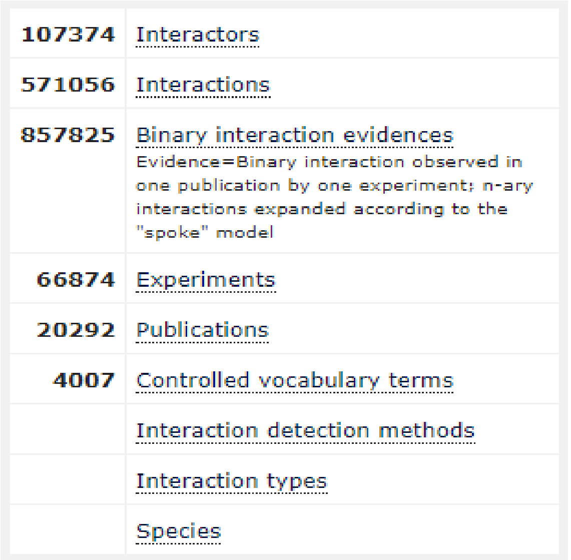
HIPPIE - Human Integrated Protein-Protein Interaction rEference



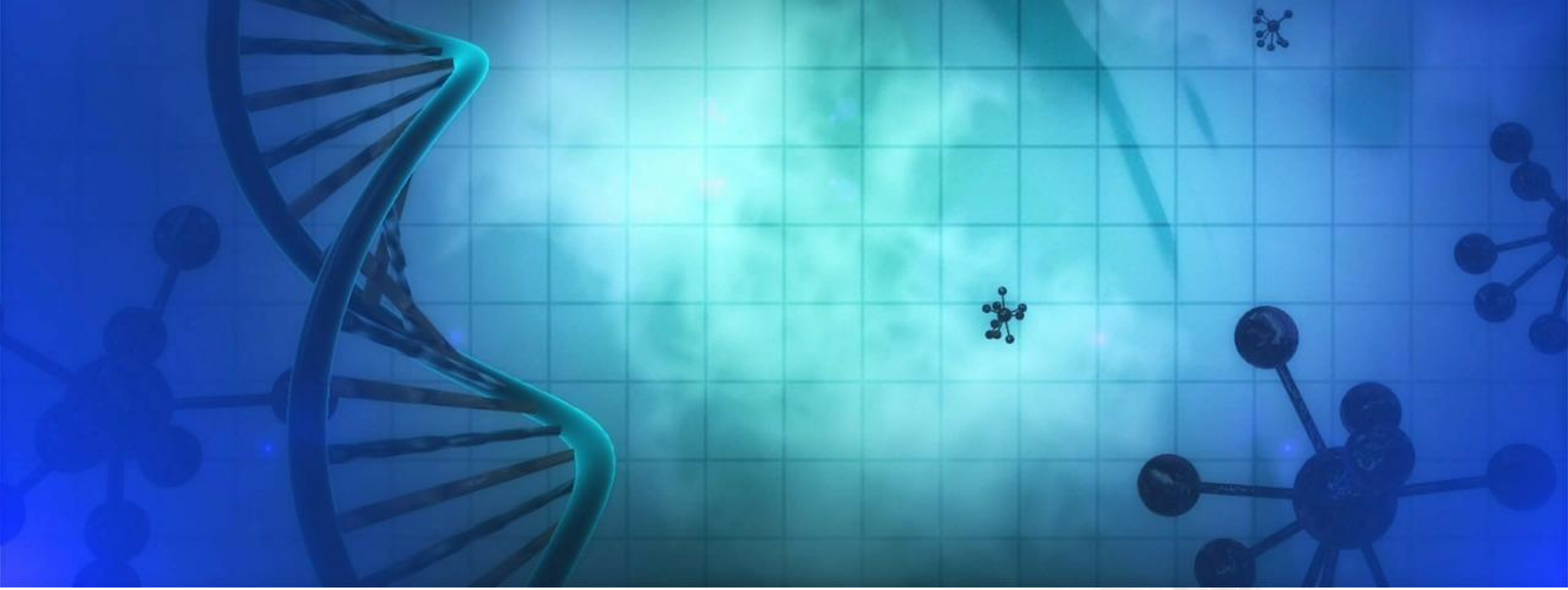
340,630 human PPIs

IntAct - database system and analysis tools for molecular interactions





107374	Interactors
571056	Interactions
857825	Binary interaction evidences Evidence=Binary interaction observed in one publication by one experiment; n-ary interactions expanded according to the "spoke" model
66874	Experiments
20292	Publications
4007	Controlled vocabulary terms
	Interaction detection methods
	Interaction types
	Species



Protein Protein Interaction (PPI) prediction problem

Importance of PPI prediction

Proteins perform their functions by interacting with other proteins

Studies:

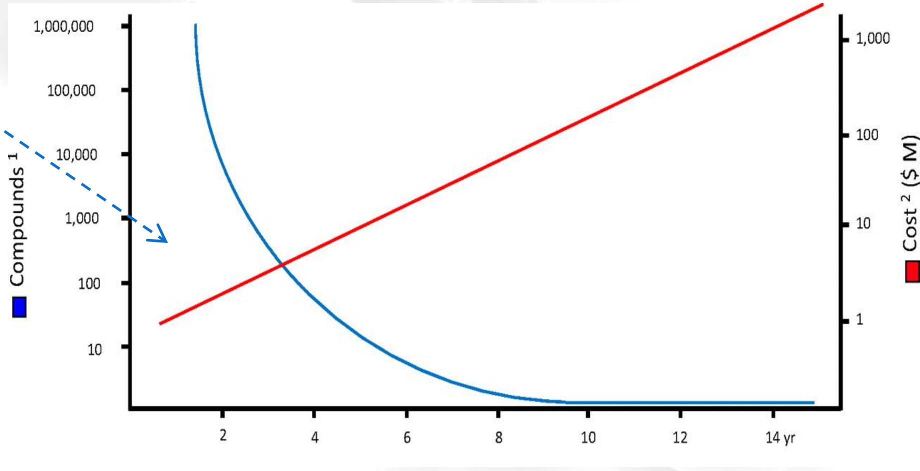
1. ***In silico*** - in computer chips
2. ***In vitro*** (in glass) – in cells, controlled env.
3. ***In vivo*** – in living organisms

Candidates

Wet lab experiments:

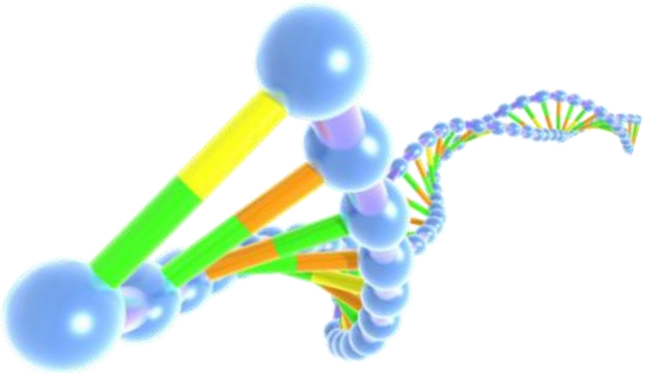
- Costly and labor-intensive
- Biases and limited coverage
- Limitations of equipment resolution
- Incomplete findings

Computer Aided Drug Discovery



Challenge of PPI prediction

>650,000 estimated Human PPIs
~340,000 human PPIs in HIPPIE DB



21,946 protein-coding human genes
240,802,485 possible human protein pairs

Complex data: PPI, co-expression, co-occurrence, GOs, Literature, Disease variants, etc.

- Heterogeneous
- Incomplete

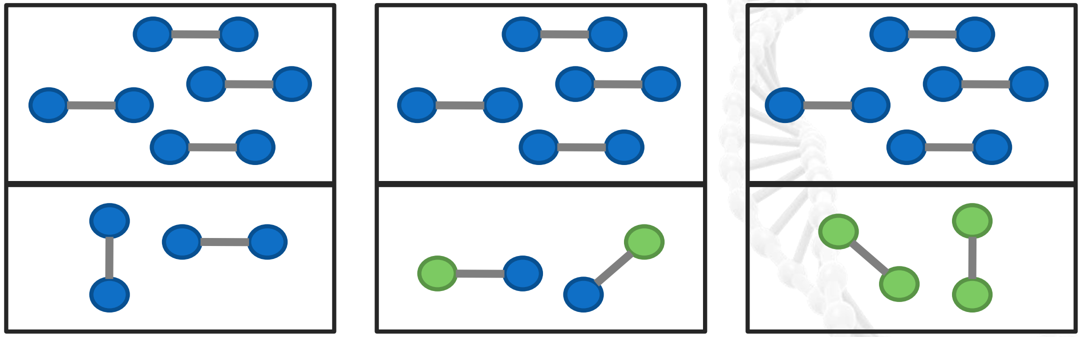
Methods based on domain knowledge => challenge

Sequence representation is universal and proteome wide available

Evaluation of PPI prediction algorithm

Three test classes of difficulties

The most difficult



TRAIN

TEST

C1

C2

C3

Accuracy

Benchmark sets *Human_Park* [Park and Marcotte, 2012]

- <40% sequence similarity
- 40 human train sets ~ **28,000** pairs
- 40 C1 test sets ~ **3,000** pairs
- 40 C2 test sets ~ **2,000** pairs
- 40 C3 test sets ~ **2,000** pairs
- Negative protein pairs were randomly sampled
- Balanced sets

Evaluation

- C1 test
- C2 test
- C3 test

Symmetric prediction

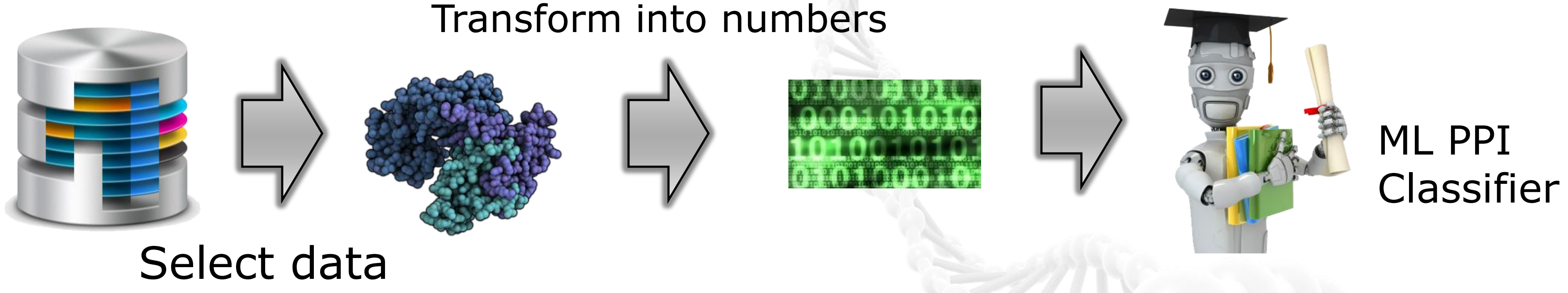
p(AB) = p(BA)

A,B proteins

Human PPI prediction
Proteome-wide approach

PPI modeling

PPI modeling process

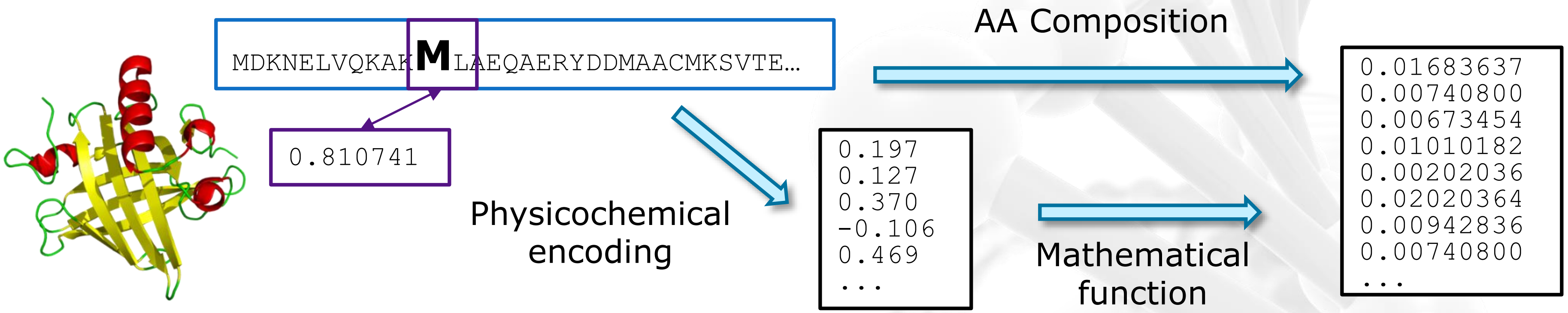


Transform into numbers

Select data

ML PPI Classifier

Coding of proteins into feature vectors



MDKNELVQKAK**M**LAEQAERYDDMAACMKSVTE…

0.810741

Physicochemical encoding

```
0.197
0.127
0.370
-0.106
0.469
...
```

AA Composition

Mathematical function

```
0.01683637
0.00740800
0.00673454
0.01010182
0.00202036
0.02020364
0.00942836
0.00740800
...
```

Coding of PPs into feature vectors

SOS1_HUMAN	GRB2_HUMAN
GRB2_HUMAN	CBL_HUMAN
MYC_HUMAN	MAX_HUMAN
JUN_HUMAN	FOS_HUMAN
RFA2_HUMAN	RFA1_HUMAN

Concatenation

0.017657	0.007270	0.006751	0.006751	....
0.013877	0.013106	0.003084	0.008866	....
0.007802	0.008173	0.008916	0.008916	....
0.012708	0.011230	0.006206	0.009161	....
0.012578	0.005296	0.007944	0.011916	....

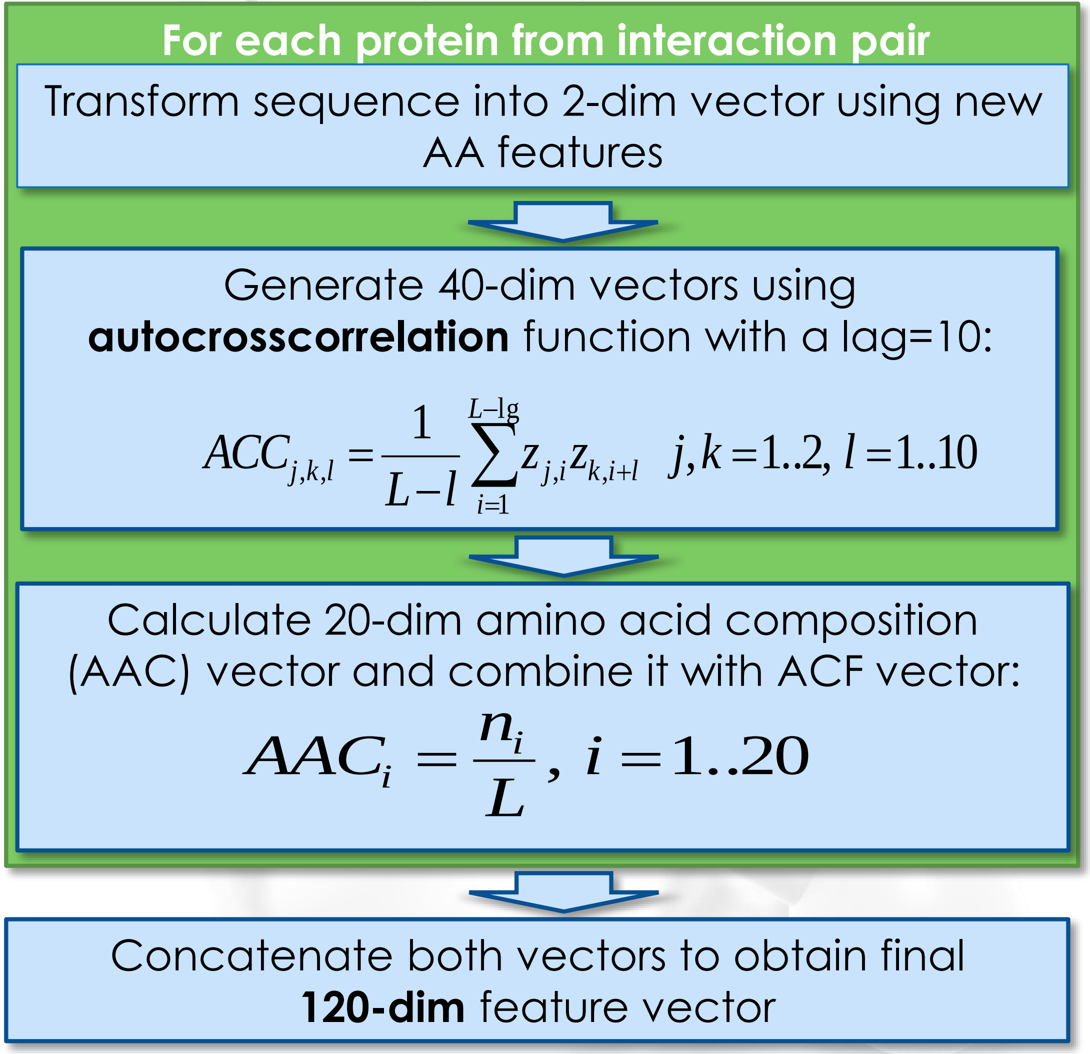
$(A,B) \in Ts$

$(B,A) \in Ts$

PCAACC Protein Encoding

Based on
- protein sequences
- amino acid physicochemical properties

Calculating PCAACC feature vector for the protein pair



For each protein from interaction pair

Transform sequence into 2-dim vector using new AA features

Generate 40-dim vectors using **autocrosscorrelation** function with a lag=10:

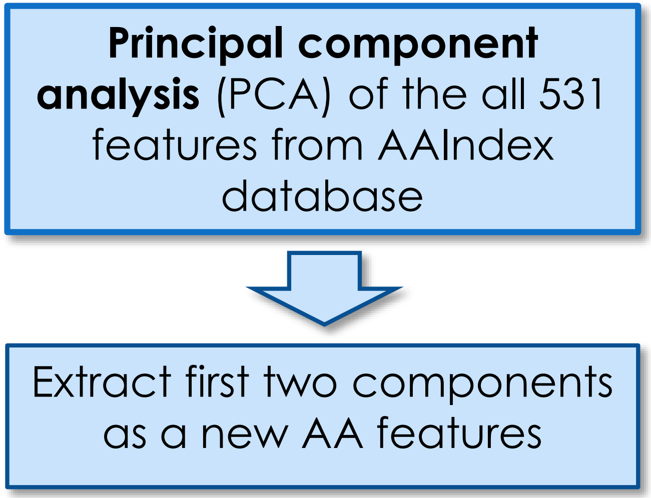
$$ACC_{j,k,l} = \frac{1}{L-l}\sum_{i=1}^{L-lg} z_{j,i} z_{k,i+l} \quad j,k = 1..2,\ l = 1..10$$

Calculate 20-dim amino acid composition (AAC) vector and combine it with ACF vector:

$$AAC_i = \frac{n_i}{L},\ i = 1..20$$

Concatenate both vectors to obtain final **120-dim** feature vector

Defining 2 new amino acid (AA) features



Principal component analysis (PCA) of the all 531 features from AAIndex database

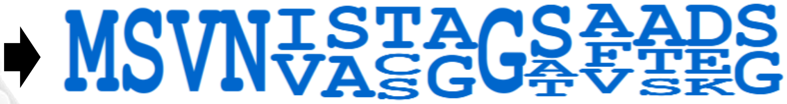
Extract first two components as a new AA features

PSSMC Protein Encoding

Position specific scoring matrix (PSSM)

- representation of evolutionary profiles using multiple sequence alignments of protein families
- determines the frequency of substitution of each amino acid at specific position in protein family - composition

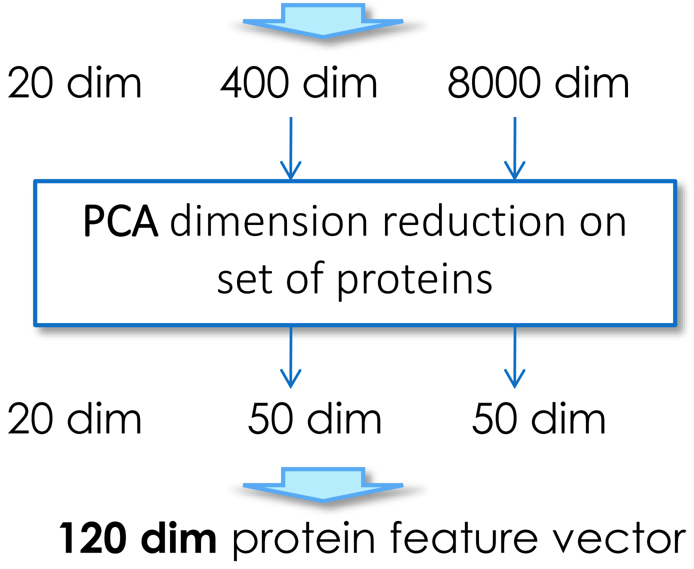
MSVNISTAGSFTES



PSSM features

$$PSSM^{AAC^k}{}_i = \frac{n_i^k}{N*M},\ i=1..20^k,\ k=1..3$$

n_i^k is number of occurrences of i-th amino acid K-tuple in N x M dimensional PSSM matrix



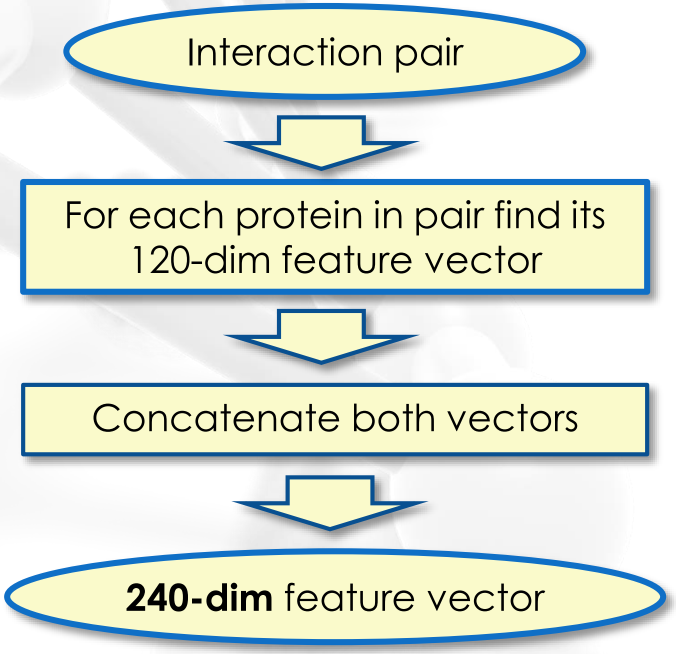
20 dim 400 dim 8000 dim

PCA dimension reduction on set of proteins

20 dim 50 dim 50 dim

120 dim protein feature vector

Generating PSSMC feature vector for protein pair



Interaction pair

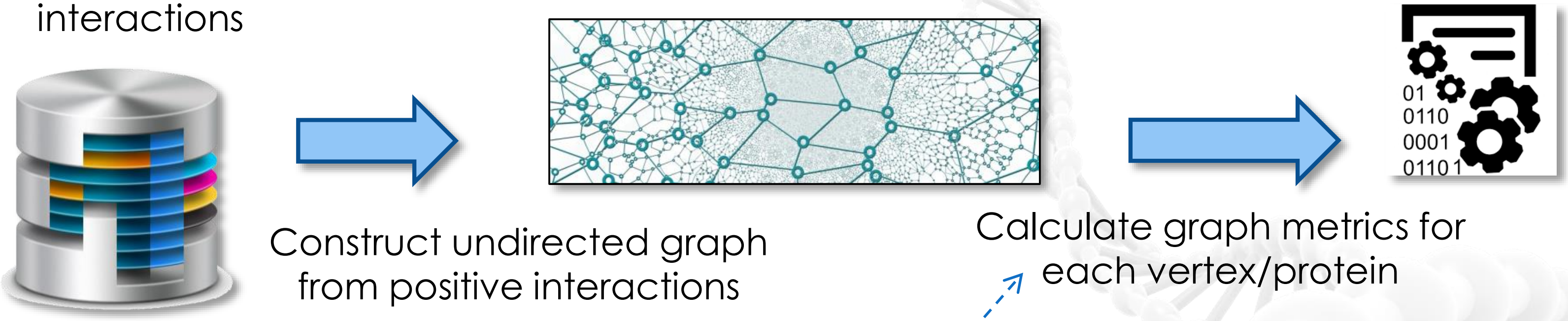
For each protein in pair find its 120-dim feature vector

Concatenate both vectors

240-dim feature vector

GraphM Protein Encoding

Calculating GraphM protein features



Training set of interactions

protein = vertex

20-dim feature vector

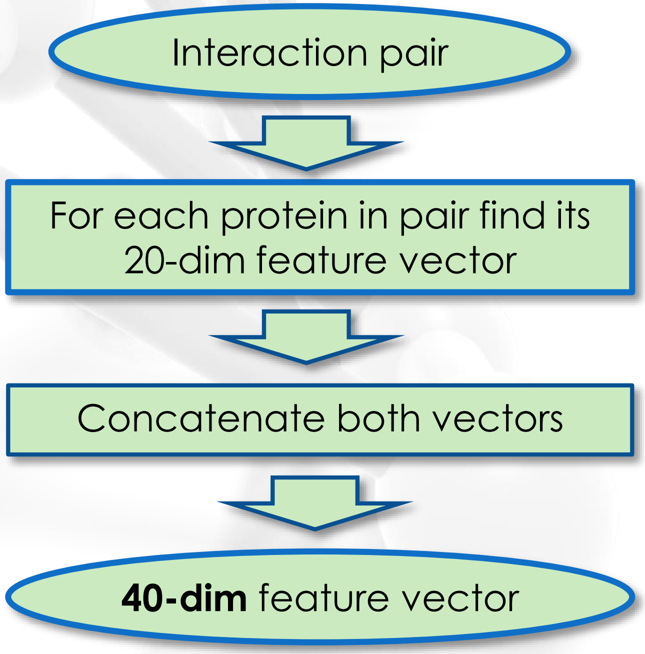
Construct undirected graph from positive interactions

Calculate graph metrics for each vertex/protein

Graph metrics used to encode the proteins

- Alpha_centrality
- Authority_score
- Betweenness
- Centrality_score
- Closeness
- Cluster_fast_greedy
- Cluster_walktrap
- Components
- Constraint
- Coreness
- Count_triangles
- Degree
- Eccentricity
- Ego
- Eigen_centrality
- Knn
- Local_scan
- Max_cardinality
- Page_rank
- Strength

Generating GraphM feature vector for protein pair (C1 class)



Interaction pair

For each protein in pair find its 20-dim feature vector

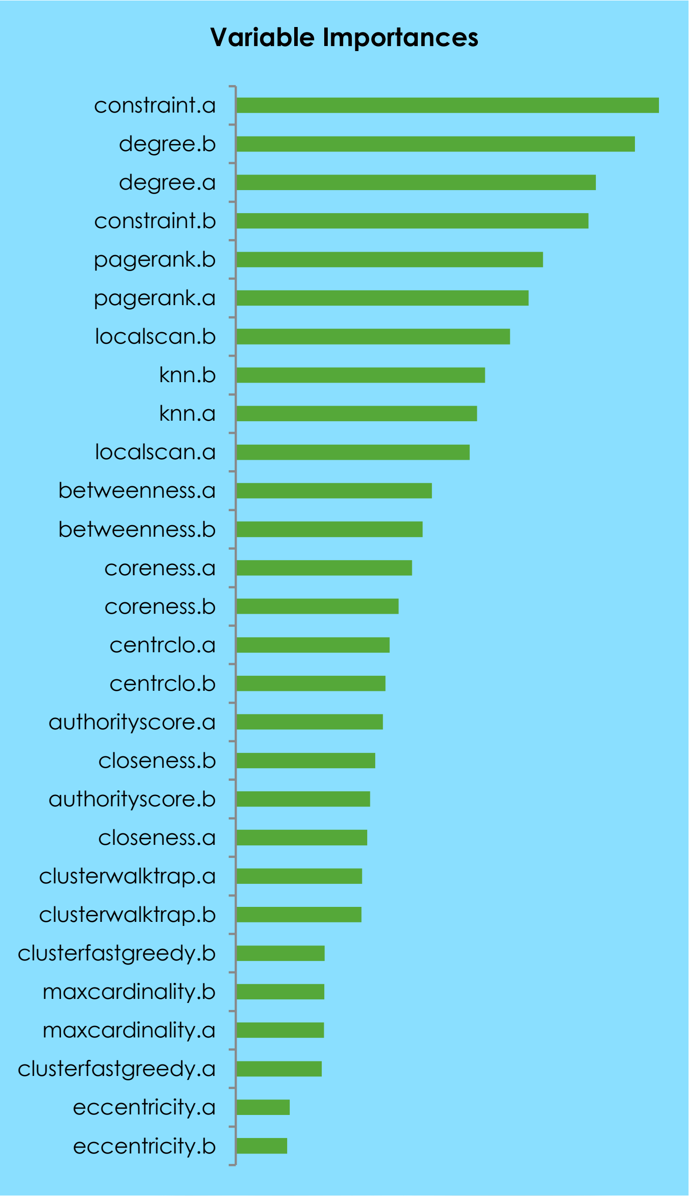
Concatenate both vectors

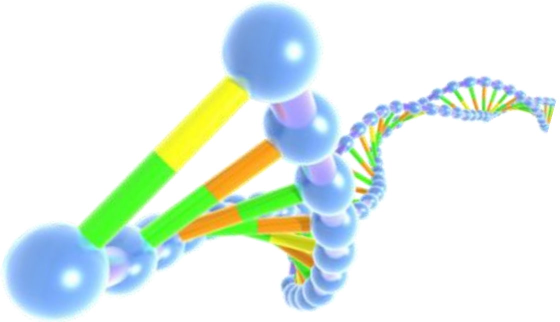
40-dim feature vector

GraphM Protein Encoding

PPI graph measures/metrics are used to encode the proteins

• **Authorityscore**	Kleinberg's authority centrality scores
• **Betweenness**	Vertex betweenness centrality
• **Centrclo**	Centrality score
• **Closeness**	Closeness centrality of vertices
• **Clusterfastgreedy**	Community structure via greedy optimization of modularity
• **Clusterwalktrap**	Community structure via short random walks
• **Constraint**	Burt's constraint
• **Coreness**	K-core decomposition of graphs
• **Degree**	Degree distribution of the vertices
• **Eccentricity**	Eccentricity of the vertices in a graph
• **Knn**	Average nearest neighbor degree
• **Localscan**	Local scan statistics
• **Maxcardinality**	Maximum cardinality search
• **Pagerank**	The Page Rank algorithm





Machine Learning

- Backward distributed feature selection driven by genetic algorithm
- Hyper parameter optimization by random/grid search
- Model selection

Algorithm

ML models		Distributed Random Forest	Gradient Boosted Machine	Generali-zed Linear Model	Deep Learning
Feature group	PCAACC	Model 1	Model 2	Model 3	Model 4
	GraphM	Model 5	Model 6	Model 7	Model 8
	PSSMC	Model 9	Model 10	Model 11	Model 12
	ALL	Model 13	Model 14	Model 15	Model 16

MuFEnsPPI final model = Ensemble of N<16 models
(**Mu**lti-**F**eature **Ens**emble PPI model)

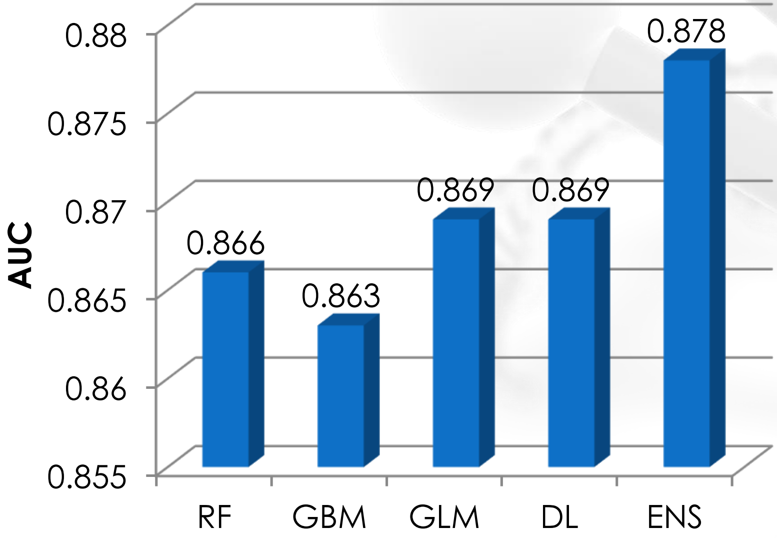
Comparison to other methods

Six state-of-the-art methods based on sequence and evolutionary profiles for PPI prediction:

- M1 [Martin et al., 2005]
- M2 [Guo et al., 2008]
- M3 [Pitre et al., 2008]
- M4 [Shen et al., 2007]
- M5 [Park & Markotte, 2012]
- M6 [Hamp & Rost, 2015]

Performance statistics on 40 YEAST C1 class and 40 HUMAN C1, C2, C3 classes test benchmark *Human_Park* sets; AUC - Area under the receiver operating characteristic curve

Method	AUC (HUMAN C1)	AUC (YEAST C1)	AUC (HUMAN C2)	AUC (HUMAN C3)
M1	0.81 ± 0.01	0.82 ± 0.01	0.61 ± 0.01	0.58 ± 0.03
M2	0.77 ± 0.01	0.76 ± 0.02	0.57 ± 0.02	0.53 ± 0.02
M3	0.77 ± 0.01	0.75 ± 0.02	0.64 ± 0.01	0.59 ± 0.02
M4	0.64 ± 0.01	0.61 ± 0.01	0.55 ± 0.01	0.50 ± 0.00
M5	0.85 ± 0.01	0.84 ± 0.01	0.60 ± 0.01	0.58 ± 0.02
M6	0.87 ± 0.01	0.87 ± 0.02	**0.69** ± 0.01	**0.67** ± 0.02
MuFEns	**0.88** ± 0.01	**0.90** ± 0.01	**0.69** ± 0.01	**0.67** ± 0.01

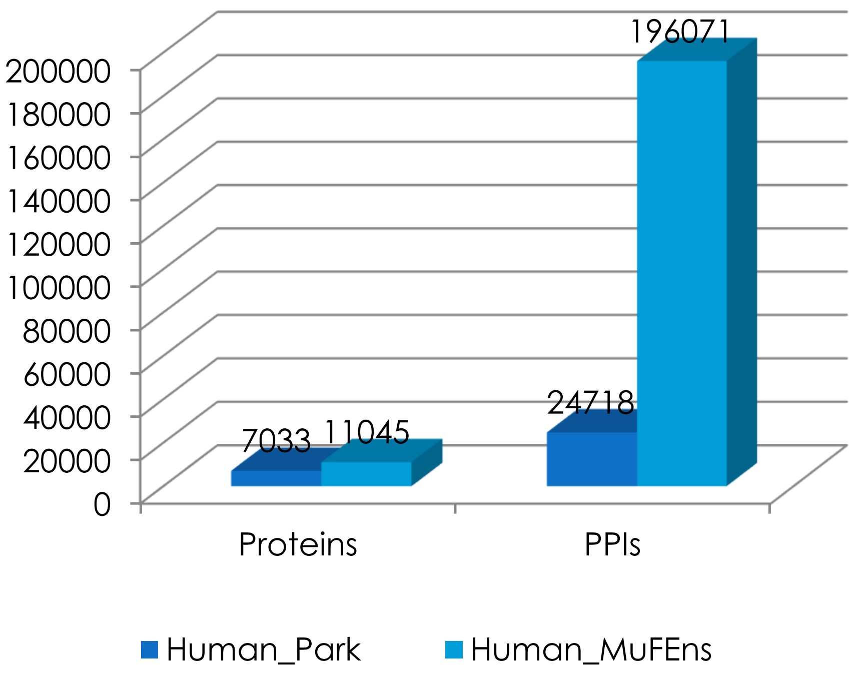


Comparison of prediction efficacy between different ML algorithms on HUMAN C1 test set using *MuFEns* model

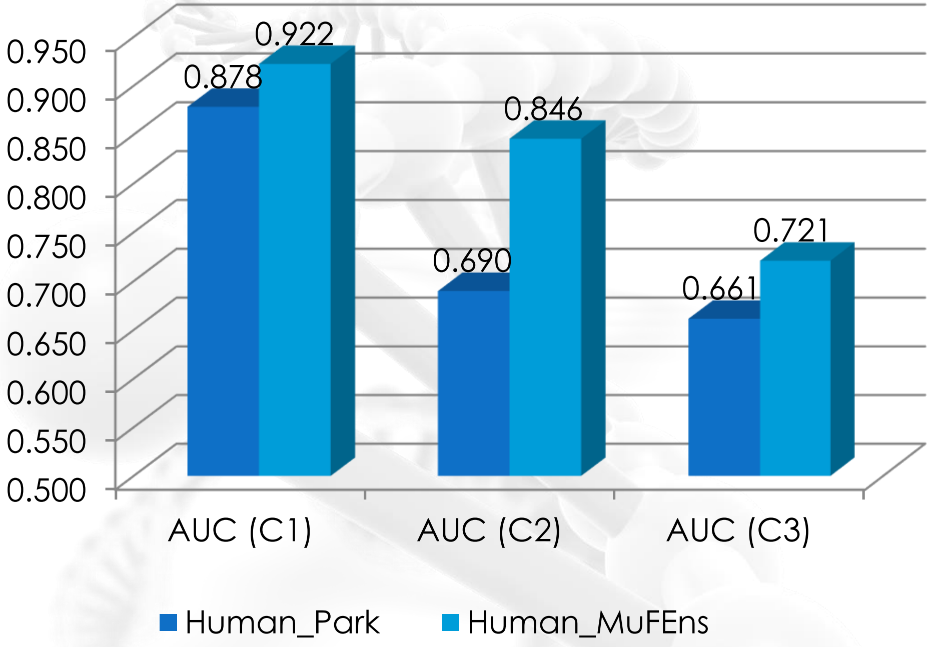
Human_MuFEns Learning Set

Human PPIs set *Human_MuFEns:* **196,000 PPIs; 11045 Proteins**

- Exclusion of >40% similar sequences and low-trust
- Negative protein pairs were randomly sampled
- Balanced sets
- 10 random splits to Train sets and C1/C2/C3 with ratio **10:1**



Increase of numbers of proteins and PPIs from *Human_Park* to *Human_MuFEns* set

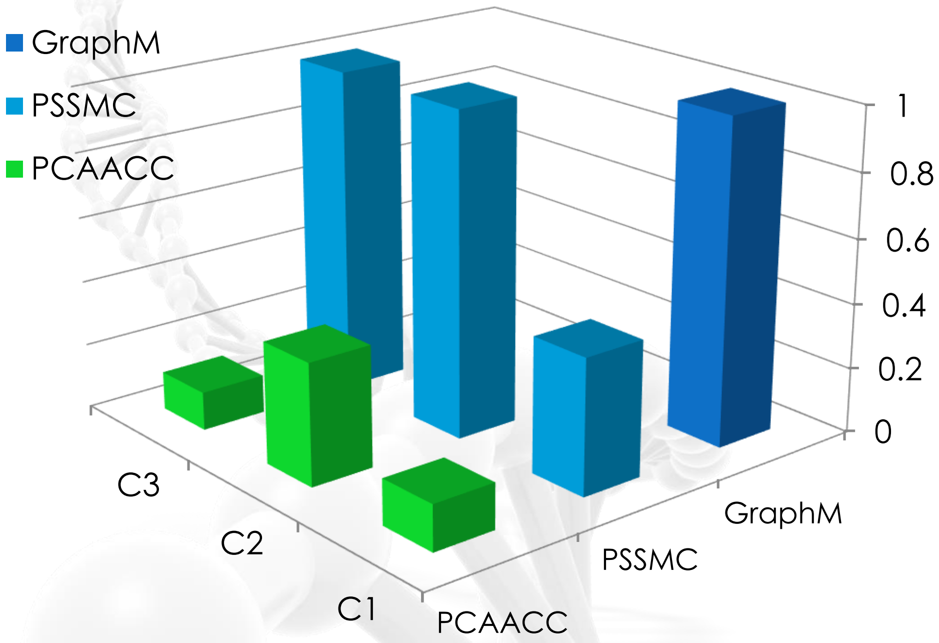


Increase of *MuFEnsPPI* model prediction performances (AUC) on new PPI test sets

Human_MuFEns Model

	C1	C2	C3
AUROC	0.922 ± 0.008	0.846 ± 0.006	0.721 ± 0.005
AUPR	0.920 ± 0.009	0.845 ± 0.007	0.643 ± 0.007
ACC	0.846 ± 0.010	0.763 ± 0.007	0.679 ± 0.006
F	0.846 ± 0.010	0.752 ± 0.008	0.716 ± 0.008
Precision	0.845 ± 0.013	0.788 ± 0.012	0.642 ± 0.011
Specificity	0.844 ± 0.018	0.807 ± 0.018	0.547 ± 0.015
Recall	0.848 ± 0.024	0.719 ± 0.020	0.810 ± 0.018
MCC	0.692 ± 0.016	0.528 ± 0.013	0.371 ± 0.012

Prediction performances of *MuFEnsPPI* model on new PPI datasets



Feature groups importances for each class

Feature calculation	
GraphM	14 min
PSSMC	4 h 20 min
PCAACC	4 min
ML training	
RF	11 min
GBM	1 h 14 min
GLM	2 min
DL	1 h 23 min

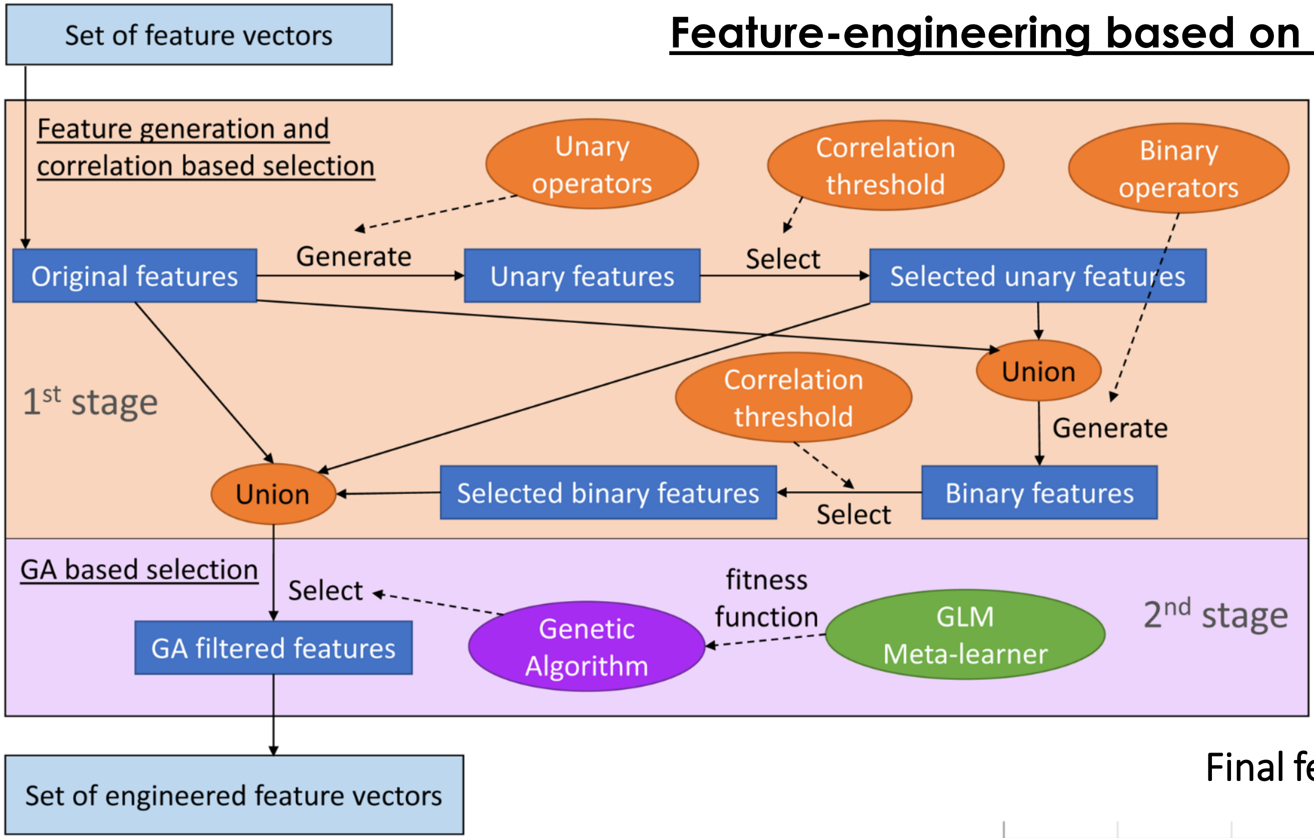
Computing times for feature calculation and ML training
Intel(R) Xeon(R) CPU E3-1230
@ 3.40GHz. 8 CPUs. 64GB RAM

HP-GAS software

Further improvement

HP-GAS Model

Feature-engineering based on Genetic Algorithm



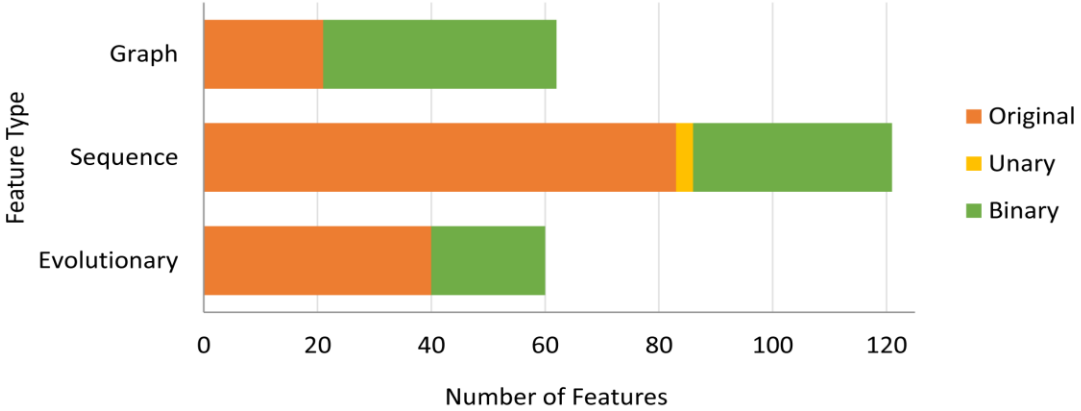
Operators:

B={sin(x), ex, 1/x, log(x), x2, √x}

U={+ , × , /}

Features	Sequence	Evolutionary	Graph
Original	120	40	42
Unary	720	240	252
Selected Unary	6	54	17
Binary	23625	13113	5133
Selected Binary	67	81	110
GA input	193	175	169
GA filtered	121	60	62

Final feature space constitution

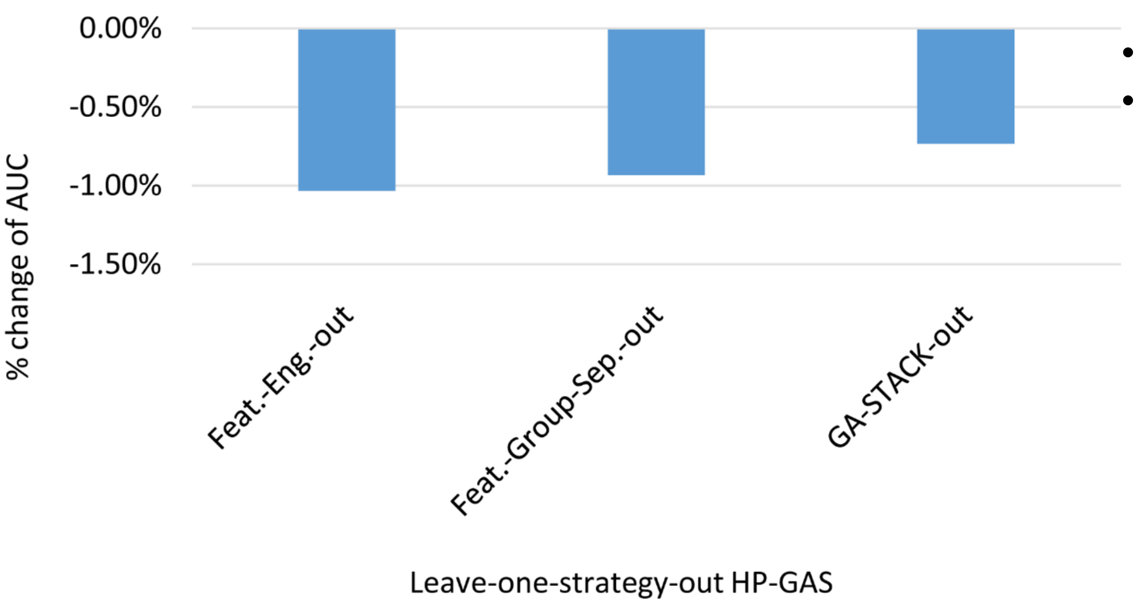


HP-GAS Model

GA-STACK ensembling algorithm based on Genetic Algorithm

- Set of base classifiers: random hyper-parameter combinations for every ML algorithm
- The fitness function of **GA** is **AUC** on the test set using training by the **GLM** supervised meta-learning algorithm which uses the predictions from models represented in individual as the features
- Crossover and mutation are bitwise operations on the '**presence**' of the models in the individual

Performances of HP-GAS in leave-one step-out experiments



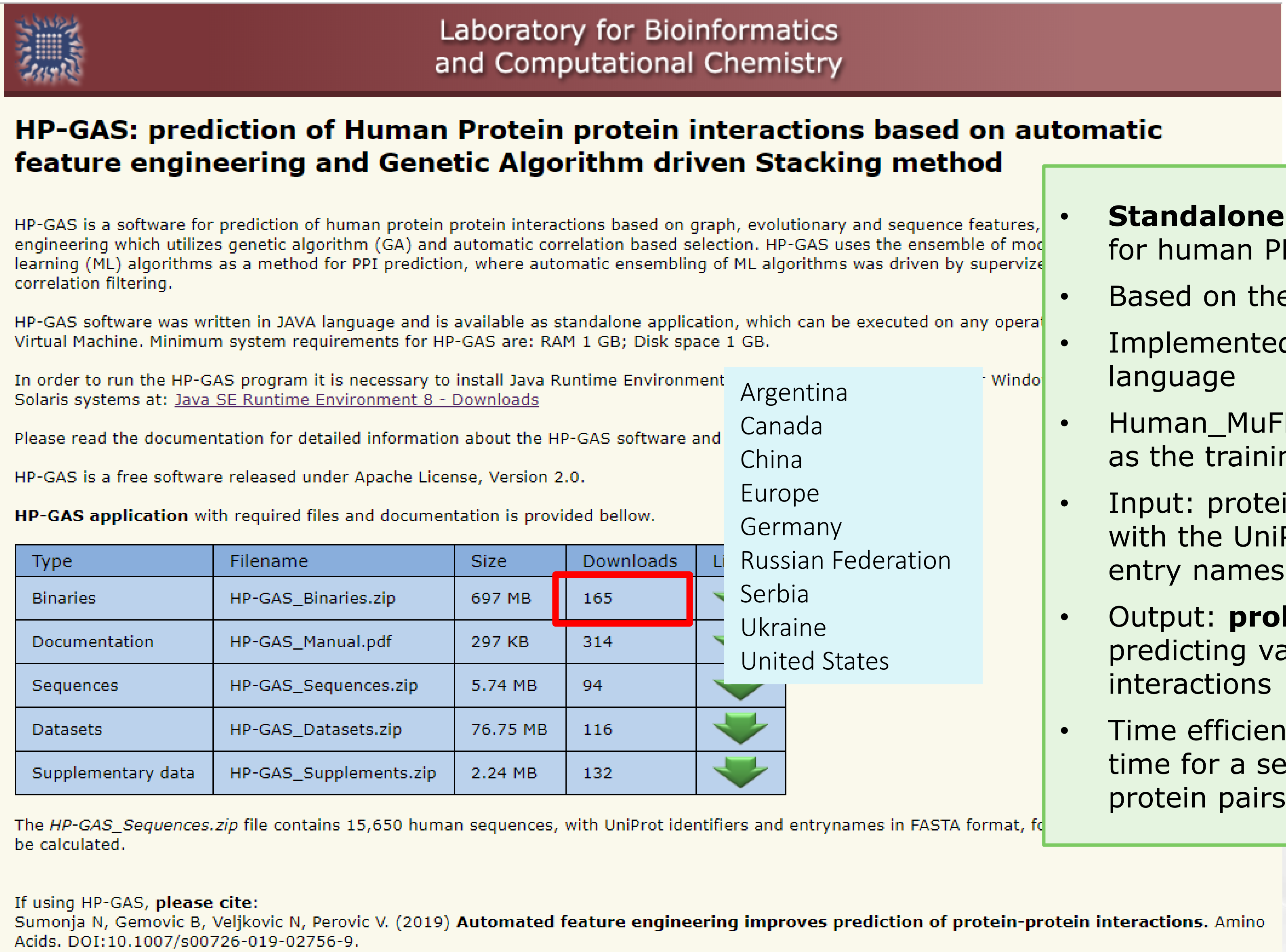
- Automatic feature generation and selection
- Feature group separation
- Meta-learner GA-STACK

Performances of HP-GAS on Human_MuFEns data sets

AUC	0.928 ± 0.001
AUPRC	0.927 ± 0.002
ACC	0.853 ± 0.002
F score	0.853 ± 0.001
MCC	0.707 ± 0.004

Sumonja N, Gemovic B, Veljkovic N, Perovic V. Automated feature engineering improves prediction of protein–protein interactions. Amino Acids. 2019; doi:10.1007/s00726-019-02756-9. ***(IF=2.5)***

HP-GAS - https://www.vin.bg.ac.rs/180/tools/HP-GAS.php



Laboratory for Bioinformatics
and Computational Chemistry

HP-GAS: prediction of Human Protein protein interactions based on automatic feature engineering and Genetic Algorithm driven Stacking method

HP-GAS is a software for prediction of human protein protein interactions based on graph, evolutionary and sequence features, engineering which utilizes genetic algorithm (GA) and automatic correlation based selection. HP-GAS uses the ensemble of mod learning (ML) algorithms as a method for PPI prediction, where automatic ensembling of ML algorithms was driven by supervize correlation filtering.

HP-GAS software was written in JAVA language and is available as standalone application, which can be executed on any operat Virtual Machine. Minimum system requirements for HP-GAS are: RAM 1 GB; Disk space 1 GB.

In order to run the HP-GAS program it is necessary to install Java Runtime Environment Windo Solaris systems at: Java SE Runtime Environment 8 - Downloads

Please read the documentation for detailed information about the HP-GAS software and

HP-GAS is a free software released under Apache License, Version 2.0.

HP-GAS application with required files and documentation is provided bellow.

Type	Filename	Size	Downloads	Li
Binaries	HP-GAS_Binaries.zip	697 MB	165	⬇
Documentation	HP-GAS_Manual.pdf	297 KB	314	⬇
Sequences	HP-GAS_Sequences.zip	5.74 MB	94	⬇
Datasets	HP-GAS_Datasets.zip	76.75 MB	116	⬇
Supplementary data	HP-GAS_Supplements.zip	2.24 MB	132	⬇

The *HP-GAS_Sequences.zip* file contains 15,650 human sequences, with UniProt identifiers and entrynames in FASTA format, fo be calculated.

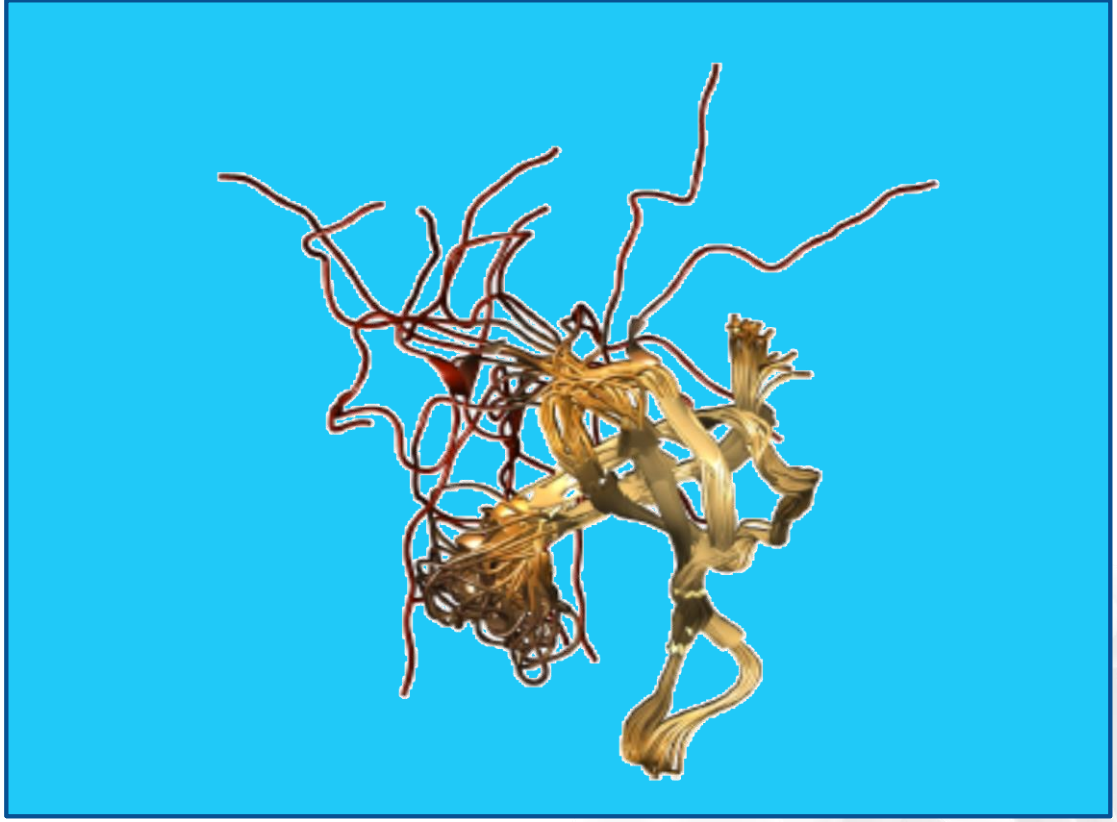
If using HP-GAS, **please cite**:
Sumonja N, Gemovic B, Veljkovic N, Perovic V. (2019) **Automated feature engineering improves prediction of protein-protein interactions.** Amino Acids. DOI:10.1007/s00726-019-02756-9.

Argentina
Canada
China
Europe
Germany
Russian Federation
Serbia
Ukraine
United States

- **Standalone software** tool for human PPI prediction
- Based on the HP-GAS model
- Implemented in **JAVA** language
- Human_MuFEns set was used as the training set
- Input: protein pairs given with the UniProt identifiers or entry names
- Output: **probabilities** as the predicting values of interactions
- Time efficient tool! Prediction time for a set of **1.000.000** protein pairs is ~**10 min**

Sumonja N, Gemovic B, Veljkovic N, Perovic V. Automated feature engineering improves prediction of protein–protein interactions. Amino Acids. 2019; doi:10.1007/s00726-019-02756-9. ***(IF=2.5)***

Human PPI prediction
Class-speciffic approach

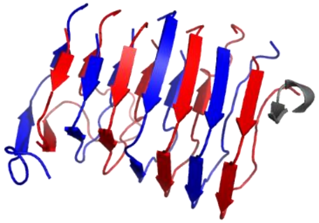


Human Intrinsically Disordered Protein Interactions prediction

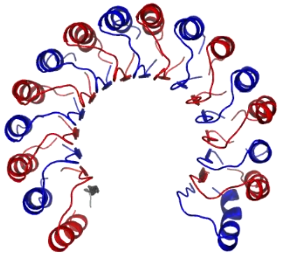
Tandem Repeat Proteins

collagen triple-helix

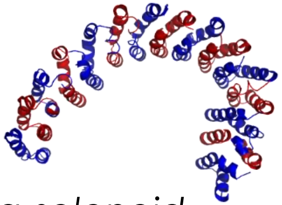
a helical coiled coil



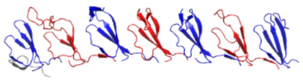
β-solenoid



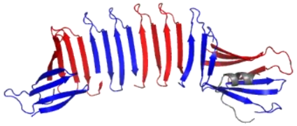
a/β solenoid



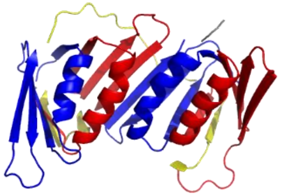
a-solenoid



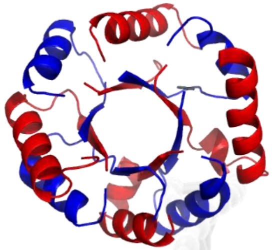
β trefoil / β hairpins



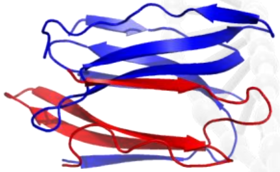
anti-parallel β layer



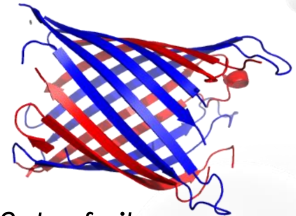
box



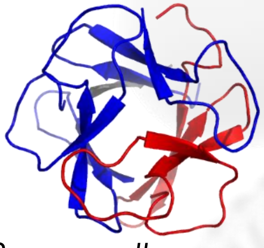
TIM-barrel



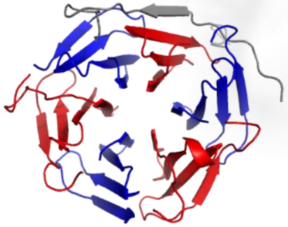
β-barrel / β hairpins



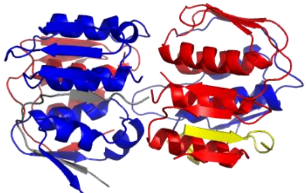
β-trefoil



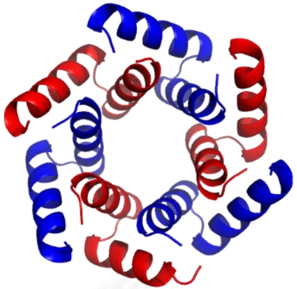
β-propeller



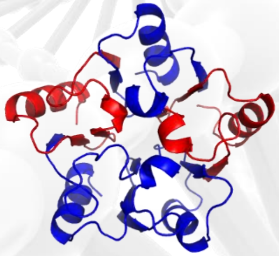
a/β prism



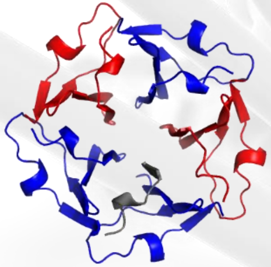
a-barrel



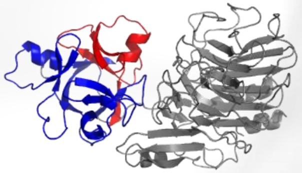
a/β barrel



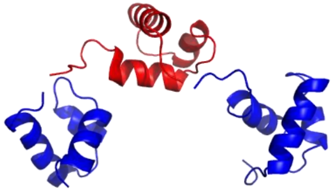
a/β propeller



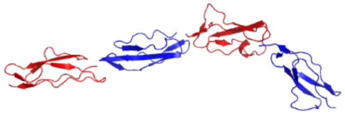
a/β trefoil



aligned prism

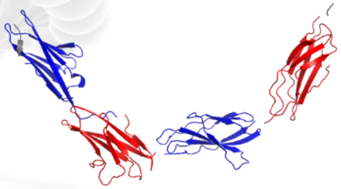


a-beads

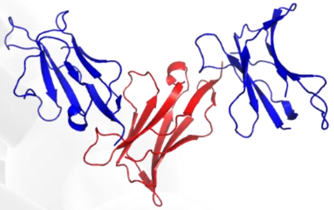


β-beads

a/β-beads



β sandwich beads



a/β sandwich beads

RepeatsDB

Paladin et al , Nucleic Acids Res. 2016

IDPpi_tool - Human Intrinsically Disordered Protein Interactions

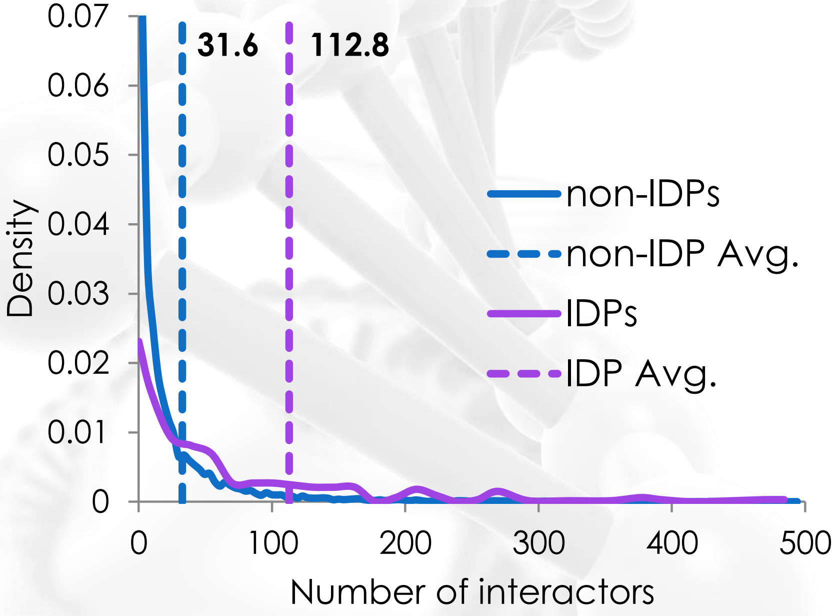
Intrinsically Disordered Proteins

- The lack of a fixed tertiary structure
- ~33% IDPs biologically functional in Eukaryota
- Biased amino acid composition and low sequence complexity
  - low proportions of bulky hydrophobic amino acids
  - high proportions of charged and hydrophilic amino acids
- Functionally important: involved in the regulation of key biological processes via binding to significantly augmented protein partners.

DisProt 7.0 (2018): database of manually curated intrinsically disordered regions:

- 803 IDP proteins
- 2167 regions
- 245 human IDPs

Piovesan et al., Nucleic Acids Res, 2017



Density curves for the interactions in the HIPPIE database

Perovic et al , Sci Rep. 2018

IDPpi_tool - Human Intrinsically Disordered Protein Interactions

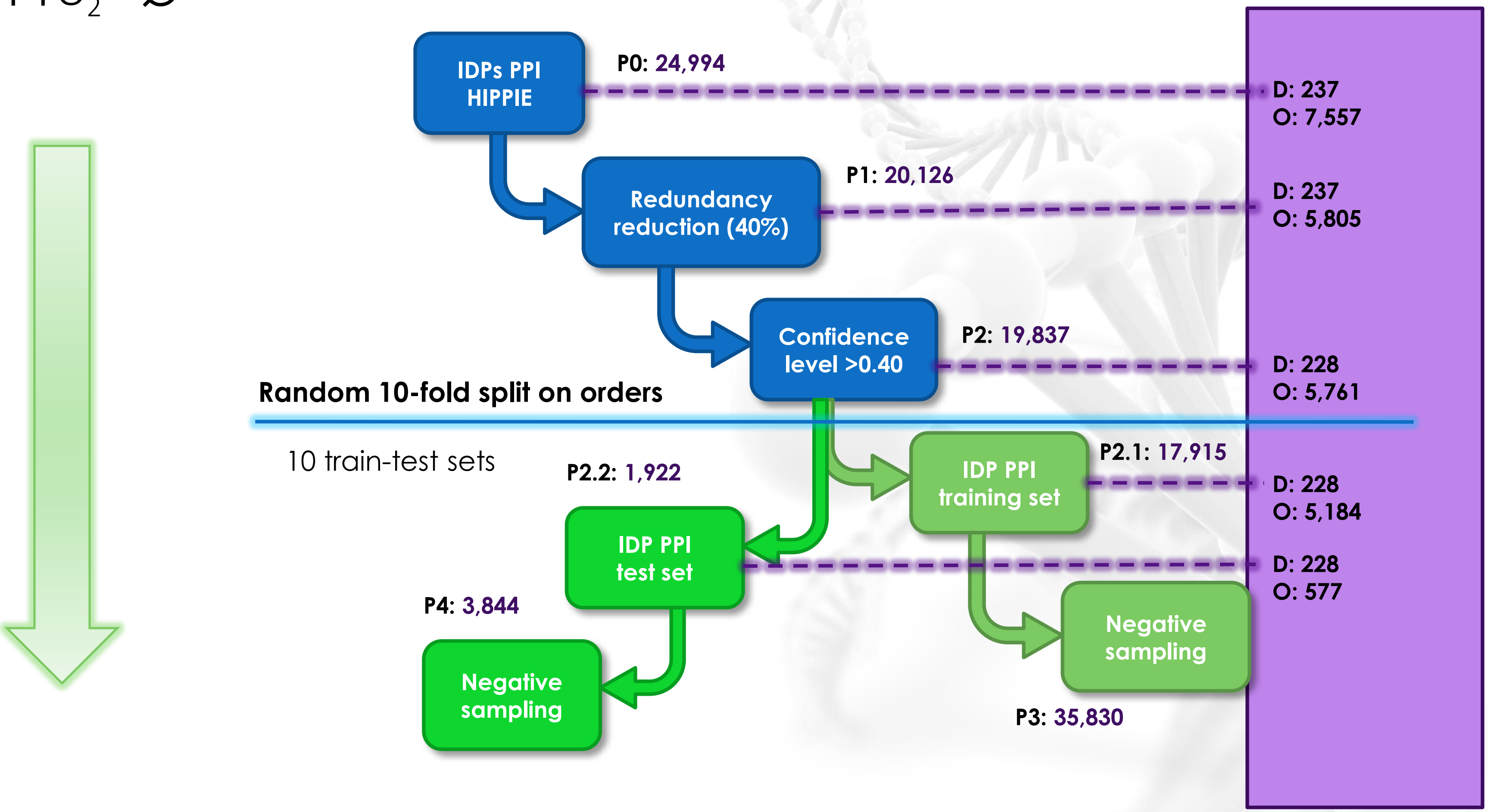
PPIs

Train (**disorder** x **order**$_1$), order$_1 \in O_1$

Test (**disorder** x **order**$_2$), order$_2 \in O_2$

$O_1 \cap O_2 = \varnothing$

Process of building data sets: train and **class C2** test



Perovic et al , Sci Rep. 2018

Pseudo amino acid composition - PseAAC

Protein: $[R_1R_2R_3\ldots R_L]$ → PseAAC vector: $(p_1, p_2, \ldots, p_{20}, p_{20+1}, \ldots, p_{20+\lambda})$

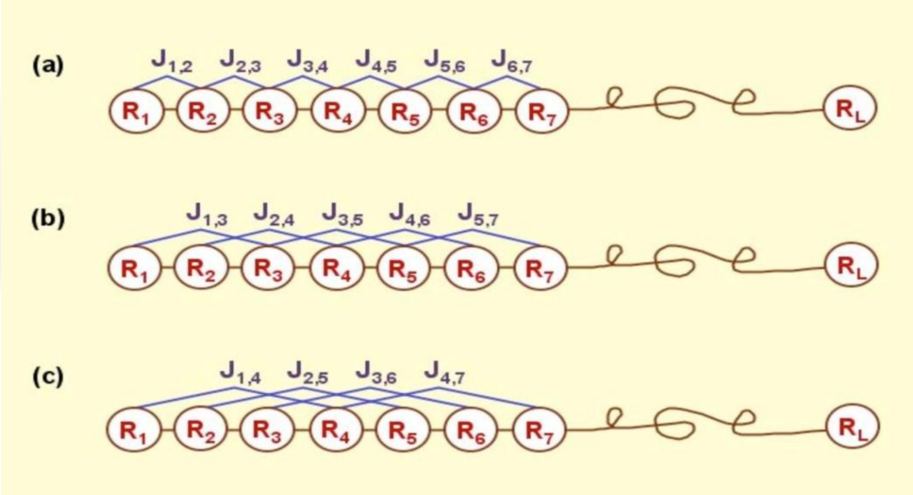
$f_1, \ldots f_{20}$ – amino acid frequencies
$\tau_1, \ldots \tau_\lambda$ – correlation coefficients $\lambda<L$

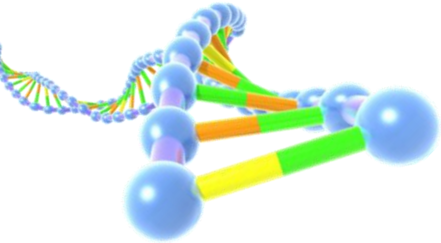
$$p_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{20} f_i + w\sum_{i=1}^{\lambda}\tau_i}, & (1 \le u \le 20) \\ \dfrac{w\tau_{u-20}}{\sum_{i=1}^{20} f_i + w\sum_{i=1}^{\lambda}\tau_i}, & (20+1 \le u \le 20+\lambda) \end{cases}$$

$$\tau_k = \frac{1}{L-k}\sum_{i=1}^{L-k} J_{i,i+k}, \quad (k<L)$$

$$J_{i,i+k} = \frac{1}{4}\sum_{q=1}^{n}\left[\phi_q(R_{i+k}) - \phi_q(R_i)\right]^2$$

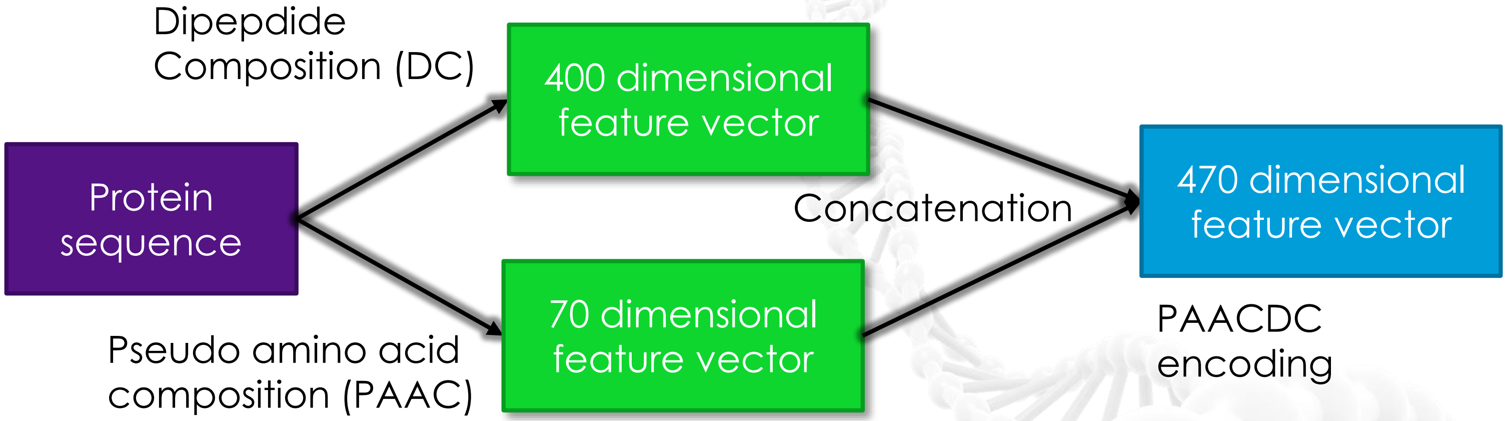
$\phi_1, \ldots, \phi_n$ - amino acid physico-chemical properties





Chou K.C.(2001). Prediction of protein cellular attributes using pseudo-amino-acid-composition. PROTEINS: Structure, Function, and Genetics 43, 246255.

IDPs representation – PAACDC features



Dipepdide Composition (DC)

400 dimensional feature vector

Protein sequence

Concatenation

470 dimensional feature vector

Pseudo amino acid composition (PAAC)

70 dimensional feature vector

PAACDC encoding

PAAC is using five disorder characteristic propensity scales:

- **TOP-IDP** scale (ranks residues by the their propensity to endorse order or disorder)
- **B-values** (flexibility parameters for each residue surrounded by two inflexible neighbours)
- **FoldUnfold** scale (capacity of amino acid residues to form a sufficient number of contacts in a globular state)
- **DisProt** scale (statistical difference in the residue compositions of ordered proteins and IDPs)
- **Net charge** scale

Method	AUC	AUPRC	ACC	F	MCC
IDPI	**0.746** ± 0.017	**0.734** ± 0.020	**0.670** ± 0.015	**0.633** ± 0.021	**0.348** ± 0.028
M1	0.688 ± 0.017	0.697 ± 0.018	0.638 ± 0.013	0.590 ± 0.022	0.285 ± 0.025
M2	0.637 ± 0.014	0.613 ± 0.012	0.593 ± 0.010	0.553 ± 0.019	0.190 ± 0.021
M3	0.627 ± 0.011	0.643 ± 0.014	0.599 ± 0.008	0.518 ± 0.013	0.211 ± 0.017

Comparison of the prediction performances between our proposed method, IDPI and other state-of-the-art sequence based methods

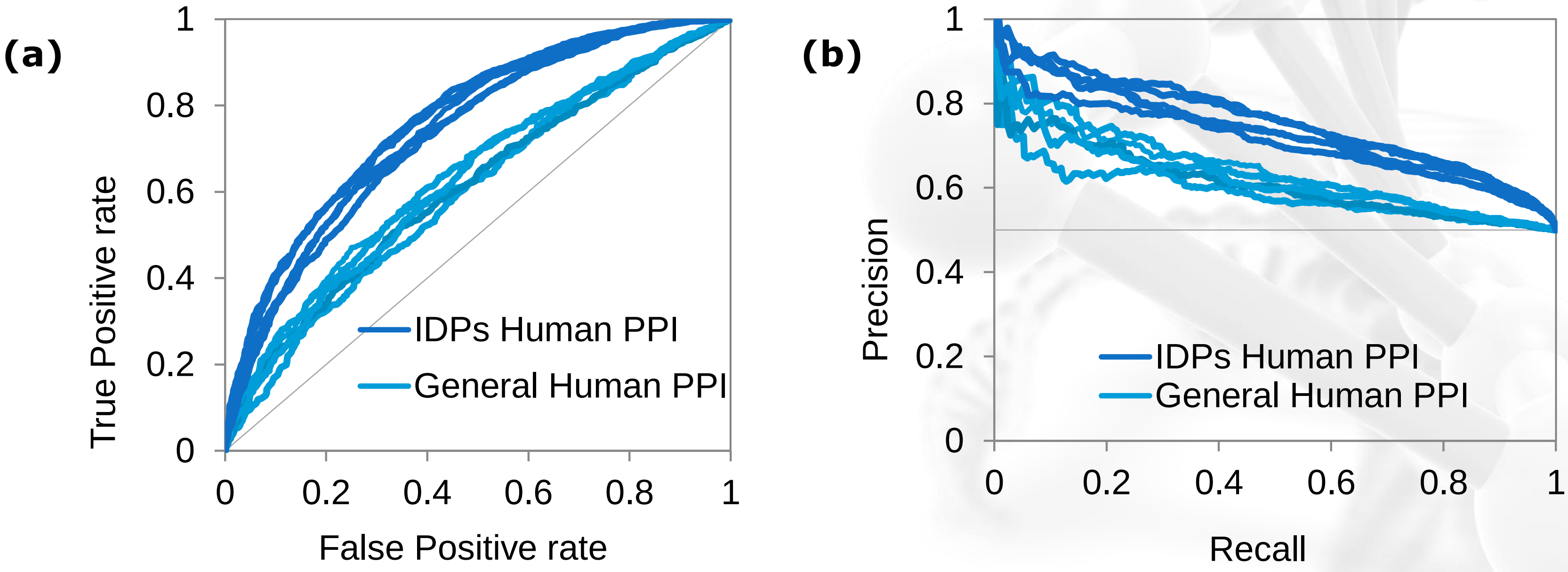
Perovic et al., Sci Rep, 2018

IDPpi_tool performances

	10N			100N		
	AUC	AUPRC	ACC	AUC	AUPRC	ACC
IDP-PPI	**0.745**	**0.237**	**0.74**	**0.748**	**0.05**	**0.757**
M1	0.691	0.217	0.724	0.692	0.048	0.737
M2	0.645	0.14	0.648	0.646	0.025	0.657
M3	0.624	0.163	0.74	0.624	0.032	0.763

Evaluation using a negative subsets randomly chosen from the negative set, where N is the size of the positive set

Comparison of predictive performances through (a) ROC curves and (b) precision/recall plots, across 5 IDP **C2** test sets using corresponding 5 IDPs and 5 general human PPI train sets.

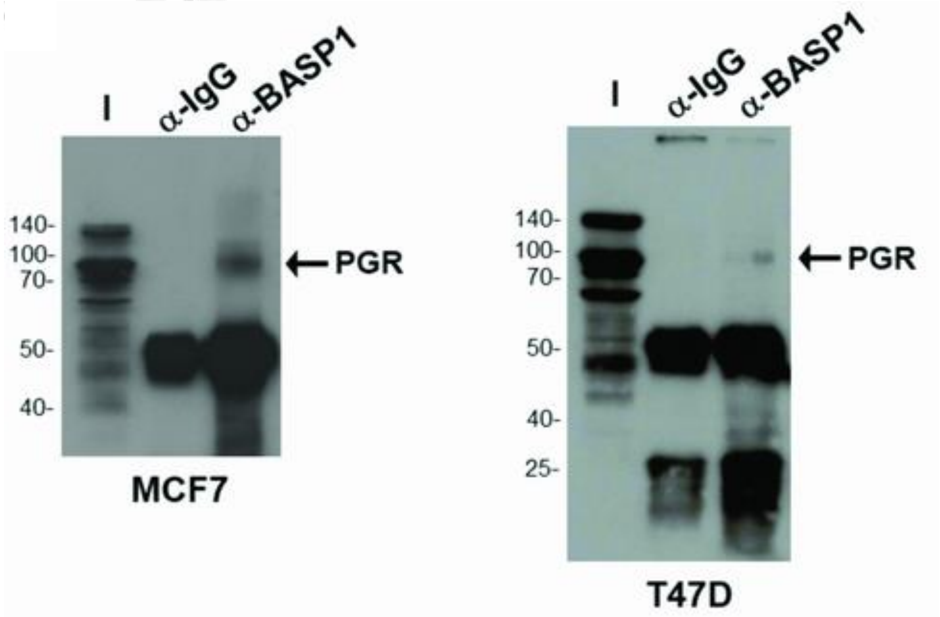


Perovic et al., Sci Rep, 2018

IDPpi_tool – new interactor identification

Example: Interactome map of Brain acid-soluble protein-1 (BASP1)

- Transcriptional cofactor
- Intrinsically disordered structure
- Silenced in several tumor types

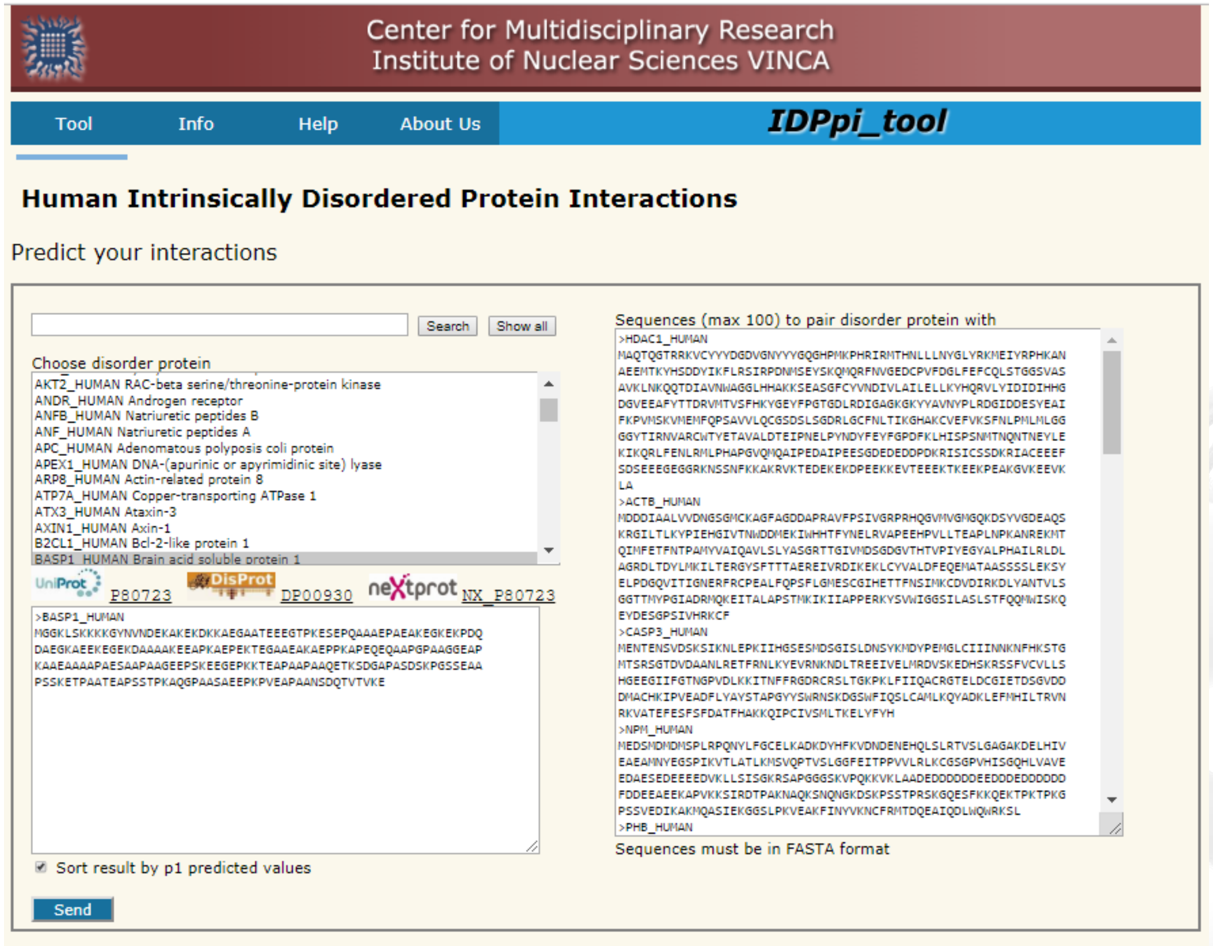


Predicted interaction between BASP1 and progesterone receptor, PRGR: In vivo binding confirmation

*Perovic V, Sumonja N, Marsh L, Radovanovic S, Vukicevic M, Roberts S, Veljkovic N. IDPpi: Protein-Protein Interaction Analyses of Human Intrinsically Disordered Proteins. Scientific Reports. 2018; doi: 10.1038/s41598-018-28815-x. **(IF=4.5)***

IDPpi_tool - http://www.vin.bg.ac.rs/180/tools/dispred.php

(a)


Center for Multidisciplinary Research
Institute of Nuclear Sciences VINCA
Tool
Info
Help
About Us
IDPpi_tool
Search
Show all
UniProt
DisProt
neXtprot
Send

Human Intrinsically Disordered Protein Interactions

Predict your interactions

Choose disorder protein

AKT2_HUMAN RAC-beta serine/threonine-protein kinase
ANDR_HUMAN Androgen receptor
ANFB_HUMAN Natriuretic peptides B
ANF_HUMAN Natriuretic peptides A
APC_HUMAN Adenomatous polyposis coli protein
APEX1_HUMAN DNA-(apurinic or apyrimidinic site) lyase
ARP8_HUMAN Actin-related protein 8
ATP7A_HUMAN Copper-transporting ATPase 1
ATX3_HUMAN Ataxin-3
AXIN1_HUMAN Axin-1
B2CL1_HUMAN Bcl-2-like protein 1
BASP1_HUMAN Brain acid soluble protein 1

P80723 DP00930 NX_P80723

```
>BASP1_HUMAN
MGGKLSKKKKGYNVNDEKAKEKDKKAEGAATEEEGTPKESEPQAAAEPAEAKEGKEKPDQ
DAEGKAEEKEGEKDAAAAKEEAPKAEPEKTEGAAEAKAEPPKAPEQEQAAPGPAAGGEAP
KAAEAAAAPAESAAPAAGEEPSKEEGEPKKTEAPAAPAAQETKSDGAPASDSKPGSSEAA
PSSKETPAATEAPSSTPKAQGPAASAEEPKPVEAPAANSDQTVTVKE
```

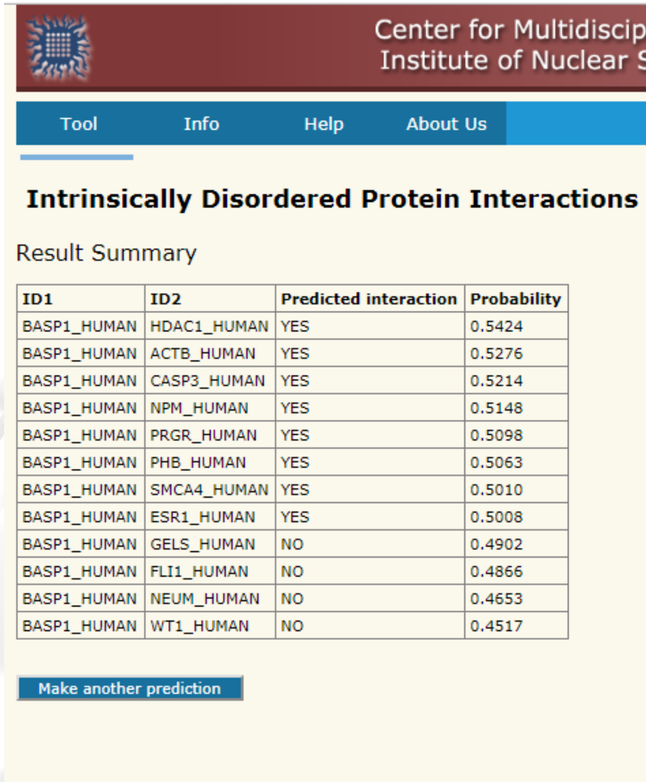
☑ Sort result by p1 predicted values

Sequences (max 100) to pair disorder protein with

```
>HDAC1_HUMAN
MAQTQGTRRKVCYYYDGDVGNYYYGQGHPMKPHRIRMTHNLLLNYGLYRKMEIYRPHKAN
AEEMTKYHSDDYIKFLRSIRPDNMSEYSKQMQRFNVGEDCPVFDGLFEFCQLSTGGSVAS
AVKLNKQQTDIAVNWAGGLHHAKKSEASGFCYVNDIVLAILELLKYHQRVLYIDIDIHHG
DGVEEAFYTTDRVMTVSFHKYGEYFPGTGDLRDIGAGKGKYYAVNYPLRDGIDDESYEAI
FKPVMSKVMEMFQPSAVVLQCGSDSLSGDRLGCFNLTIKGHAKCVEFVKSFNLPMLMLGG
GGYTIRNVARCWTYETAVALDTEIPNELPYNDYFEYFGPDFKLHISPSNMTNQNTNEYLE
KIKQRLFENLRMLPHAPGVQMQAIPEDAIPEESGDEDEDDPDKRISICSSDKRIACEEEF
SDSEEEGEGGRKNSSNFKKAKRVKTEDEKEKDPEEKKEVTEEEKTKEEKPEAKGVKEEVK
LA
>ACTB_HUMAN
MDDDIAALVVDNGSGMCKAGFAGDDAPRAVFPSIVGRPRHQGVMVGMGQKDSYVGDEAQS
KRGILTLKYPIEHGIVTNWDDMEKIWHHTFYNELRVAPEEHPVLLTEAPLNPKANREKMT
QIMFETFNTPAMYVAIQAVLSLYASGRTTGIVMDSGDGVTHTVPIYEGYALPHAILRLDL
AGRDLTDYLMKILTERGYSFTTTAEREIVRDIKEKLCYVALDFEQEMATAASSSSLEKSY
ELPDGQVITIGNERFRCPEALFQPSFLGMESCGIHETTFNSIMKCDVDIRKDLYANTVLS
GGTTMYPGIADRMQKEITALAPSTMKIKIIAPPERKYSVWIGGSILASLSTFQQMWISKQ
EYDESGPSIVHRKCF
>CASP3_HUMAN
MENTENSVDSKSIKNLEPKIIHGSESMDSGISLDNSYKMDYPEMGLCIIINNKNFHKSTG
MTSRSGTDVDAANLRETFRNLKYEVRNKNDLTREEIVELMRDVSKEDHSKRSSFVCVLLS
HGEEGIIFGTNGPVDLKKITNFFRGDRCRSLTGKPKLFIIQACRGTELDCGIETDSGVDD
DMACHKIPVEADFLYAYSTAPGYYSWRNSKDGSWFIQSLCAMLKQYADKLEFMHILTRVN
RKVATEFESFSFDATFHAKKQIPCIVSMLTKELYFYH
>NPM_HUMAN
MEDSMDMDMSPLRPQNYLFGCELKADKDYHFKVDNDENEHQLSLRTVSLGAGAKDELHIV
EAEAMNYEGSPIKVTLATLKMSVQPTVSLGGFEITPPVVLRLKCGSGPVHISGQHLVAVE
EDAESEDEEEEDVKLLSISGKRSAPGGGSKVPQKKVKLAADEDDDDDDEEDDDEDDDDDD
FDDEEAEEKAPVKKSIRDTPAKNAQKSNQNGKDSKPSSTPRSKGQESFKKQEKTPKTPKG
PSSVEDIKAKMQASIEKGGSLPKVEAKFINYVKNCFRMTDQEAIQDLWQWRKSL
>PHB_HUMAN
```

Sequences must be in FASTA format

(b)



Intrinsically Disordered Protein Interactions

Result Summary

ID1	ID2	Predicted interaction	Probability
BASP1_HUMAN	HDAC1_HUMAN	YES	0.5424
BASP1_HUMAN	ACTB_HUMAN	YES	0.5276
BASP1_HUMAN	CASP3_HUMAN	YES	0.5214
BASP1_HUMAN	NPM_HUMAN	YES	0.5148
BASP1_HUMAN	PRGR_HUMAN	YES	0.5098
BASP1_HUMAN	PHB_HUMAN	YES	0.5063
BASP1_HUMAN	SMCA4_HUMAN	YES	0.5010
BASP1_HUMAN	ESR1_HUMAN	YES	0.5008
BASP1_HUMAN	GELS_HUMAN	NO	0.4902
BASP1_HUMAN	FLI1_HUMAN	NO	0.4866
BASP1_HUMAN	NEUM_HUMAN	NO	0.4653
BASP1_HUMAN	WT1_HUMAN	NO	0.4517

Time efficient tool!
Prediction time for
100 protein pairs is
less than a second

IDPpi_tool Web Interface (a) Front page of *IDPpi_tool* web application where users can input the protein sequences in a FASTA format and to choose either automatic combination in pairs or to add protein pairs of interest to the input information. (b) *IDPpi_tool* results page.

Prediction of Transcriptional Regulation Interactions

TRI_tool Prediction of Transcriptional Regulation Interactions

Transcriptional regulation (TR) is a complex process which controls the cellular gene expression and among the key processes in all serious human diseases, including cancer.
It is important to identify pharmacologically relevant PPIs.

Datasets and models

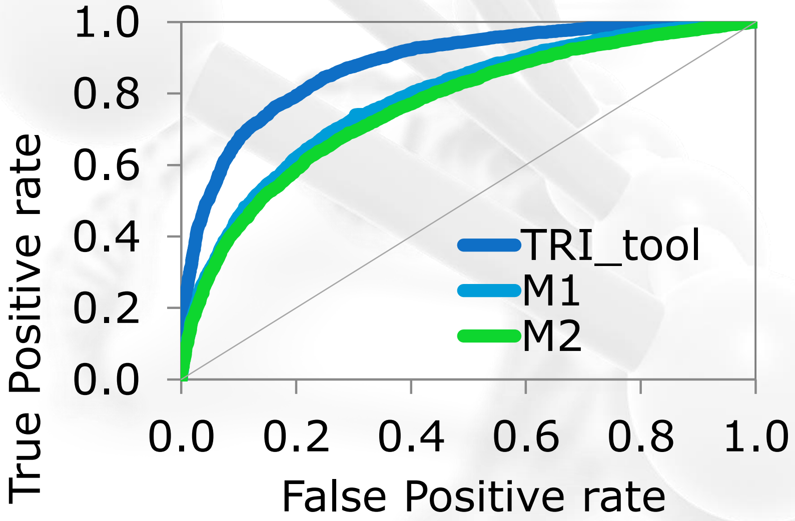
1515 proteins involved in human transcriptional regulation (UniProt)
12244 mutual interactions (HIPPIE - Human Integrated Protein-Protein Interaction rEference)

Performances in prediction efficiency

Comparison between TRI_tool and two state-of-the-art sequence-based methods:

M1 (Guo et al., 2008)

M2 (Pitre et al., 2008)



Perovic et al., Bioinformatics, 2017

Prediction of Transcriptional Regulation Interactions

TRI_tool – web service

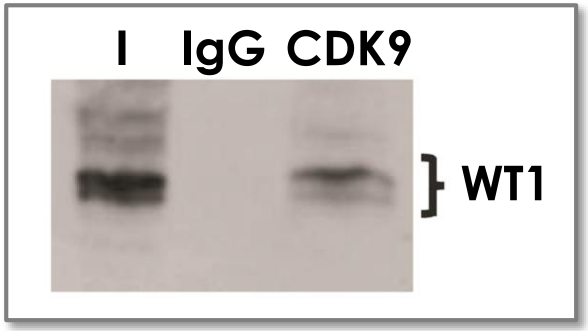
http://www.vin.bg.ac.rs/180/tools/tfpred.php

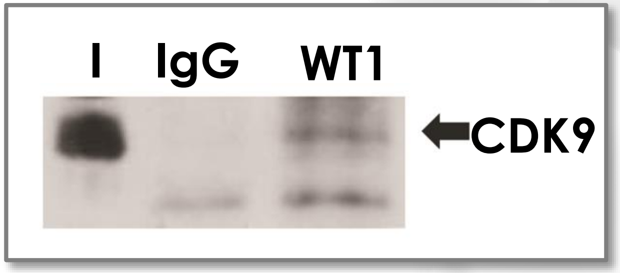
Effective in dealing with **large number of sequences** and outperforms some of the mostly used sequence-based methods in terms of computational efficacy and prediction potential.

- 100 interactions in less then a second!

TRI_tool predicted WT1-CDK9 interaction

Identification of a new interacting partner for Wilm's tumor protein (**WT1**): Anti-cancer target cyclin-dependent kinase (**CDK9**)

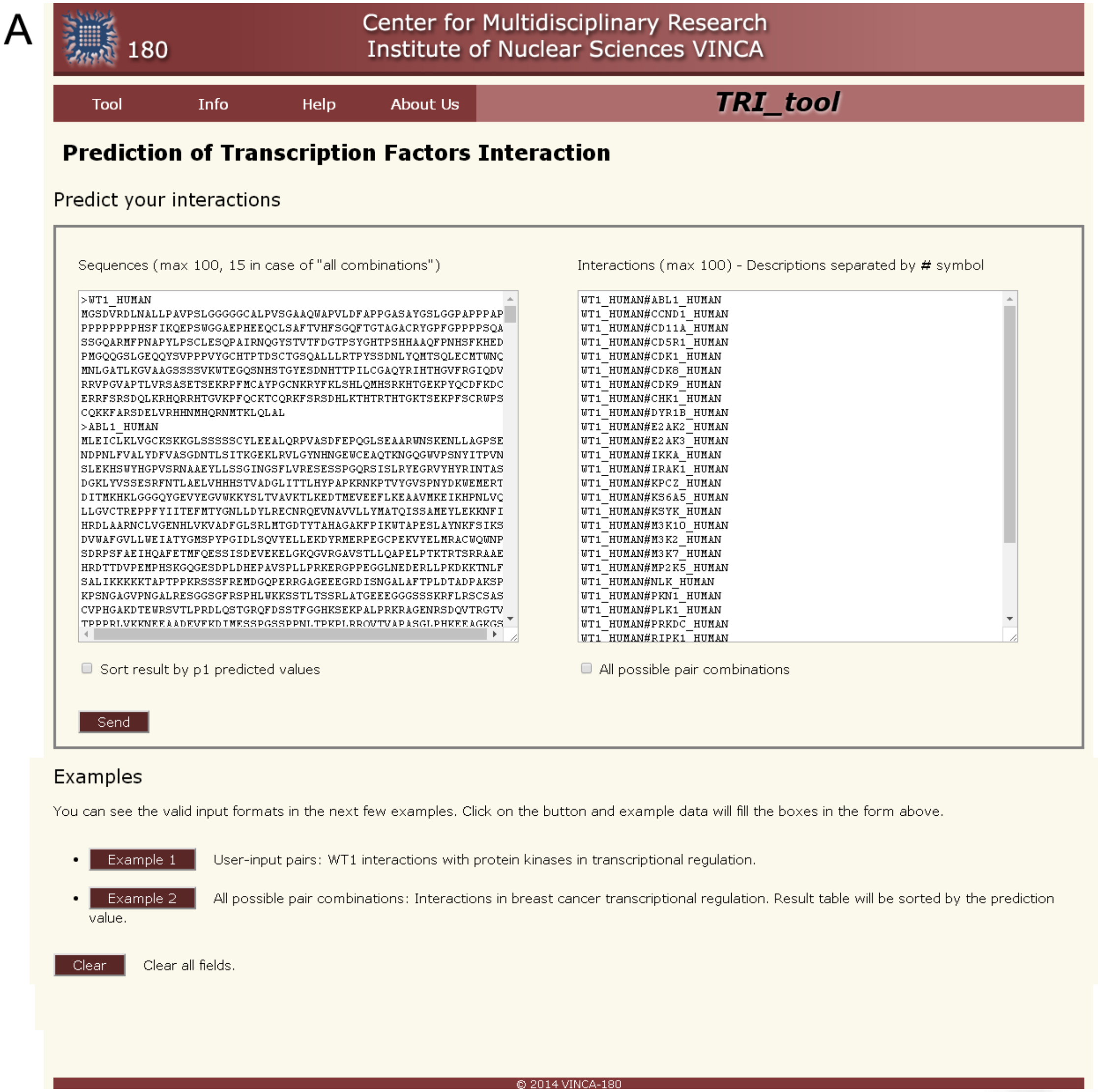




In vivo binding confirmation. Co-immunoprecipitation of WT1 and CDK9 in human leukemia cell line K562

Perovic V, Sumonja N, Gemovic B, Toska E, Roberts SG, Veljkovic N. TRI_tool: a web-tool for prediction of protein-protein interactions in human transcriptional regulation. Bioinformatics. 2017; 33(2):289-91. (**IF=4.5**)

TRI_tool - http://www.vin.bg.ac.rs/180/tools/tfpred.php


A
180
Center for Multidisciplinary Research
Institute of Nuclear Sciences VINCA
Tool
Info
Help
About Us
TRI_tool

Prediction of Transcription Factors Interaction

Predict your interactions

Sequences (max 100, 15 in case of "all combinations")

```
>WT1_HUMAN
MGSDVRDLNALLPAVPSLGGGGGCALPVSGAAQWAPVLDFAPPGASAYGSLGGPAPPPAP
PPPPPPPPHSFIKQEPSWGGAEPHEEQCLSAFTVHFSGQFTGTAGACRYGPFGPPPPSQA
SSGQARMFPNAPYLPSCLESQPAIRNQGYSTVTFDGTPSYGHTPSHHAAQFPNHSFKHED
PMGQQGSLGEQQYSVPPPVYGCHTPTDSCTGSQALLLRTPYSSDNLYQMTSQLECMTWNQ
MNLGATLKGVAAGSSSSVKWTEGQSNHSTGYESDNHTTPILCGAQYRIHTHGVFRGIQDV
RRVPGVAPTLVRSASETSEKRPFMCAYPGCNKRYFKLSHLQMHSRKHTGEKPYQCDFKDC
ERRFSRSDQLKRHQRRHTGVKPFQCKTCQRKFSRSDHLKTHTRTHTGKTSEKPFSCRWPS
CQKKFARSDELVRHHNMHQRNMTKLQLAL
>ABL1_HUMAN
MLEICLKLVGCKSKKGLSSSSSCYLEEALQRPVASDFEPQGLSEAARWNSKENLLAGPSE
NDPNLFVALYDFVASGDNTLSITKGEKLRVLGYNHNGEWCEAQTKNGQGWVPSNYITPVN
SLEKHSWYHGPVSRNAAEYLLSSGINGSFLVRESESSPGQRSISLRYEGRVYHYRINTAS
DGKLYVSSESRFNTLAELVHHHSTVADGLITTLHYPAPKRNKPTVYGVSPNYDKWEMERT
DITMKHKLGGGQYGEVYEGVWKKYSLTVAVKTLKEDTMEVEEFLKEAAVMKEIKHPNLVQ
LLGVCTREPPFYIITEFMTYGNLLDYLRECNRQEVNAVVLLYMATQISSAMEYLEKKNFI
HRDLAARNCLVGENHLVKVADFGLSRLMTGDTYTAHAGAKFPIKWTAPESLAYNKFSIKS
DVWAFGVLLWEIATYGMSPYPGIDLSQVYELLEKDYRMERPEGCPEKVYELMRACWQWNP
SDRPSFAEIHQAFETMFQESSISDEVEKELGKQGVRGAVSTLLQAPELPTKTRTSRRAAE
HRDTTDVPEMPHSKGQGESDPLDHEPAVSPLLPRKERGPPEGGLNEDERLLPKDKKTNLF
SALIKKKKKTAPTPPKRSSSFREMDGQPERRGAGEEEGRDISNGALAFTPLDTADPAKSP
KPSNGAGVPNGALRESGGSGFRSPHLWKKSSTLTSSRLATGEEEGGGSSSKRFLRSCSAS
CVPHGAKDTEWRSVTLPRDLQSTGRQFDSSTFGGHKSEKPALPRKRAGENRSDQVTRGTV
TPPPRLVKKNEEAADEVFKDIMESSPGSSPPNLTPKPLRRQVTAPASGLPHKEEAGKGS
```

Interactions (max 100) - Descriptions separated by # symbol

```
WT1_HUMAN#ABL1_HUMAN
WT1_HUMAN#CCND1_HUMAN
WT1_HUMAN#CD11A_HUMAN
WT1_HUMAN#CD5R1_HUMAN
WT1_HUMAN#CDK1_HUMAN
WT1_HUMAN#CDK8_HUMAN
WT1_HUMAN#CDK9_HUMAN
WT1_HUMAN#CHK1_HUMAN
WT1_HUMAN#DYR1B_HUMAN
WT1_HUMAN#E2AK2_HUMAN
WT1_HUMAN#E2AK3_HUMAN
WT1_HUMAN#IKKA_HUMAN
WT1_HUMAN#IRAK1_HUMAN
WT1_HUMAN#KPCZ_HUMAN
WT1_HUMAN#KS6A5_HUMAN
WT1_HUMAN#KSYK_HUMAN
WT1_HUMAN#M3K10_HUMAN
WT1_HUMAN#M3K2_HUMAN
WT1_HUMAN#M3K7_HUMAN
WT1_HUMAN#MP2K5_HUMAN
WT1_HUMAN#NLK_HUMAN
WT1_HUMAN#PKN1_HUMAN
WT1_HUMAN#PLK1_HUMAN
WT1_HUMAN#PRKDC_HUMAN
WT1_HUMAN#RIPK1_HUMAN
```

☐ Sort result by p1 predicted values

☐ All possible pair combinations

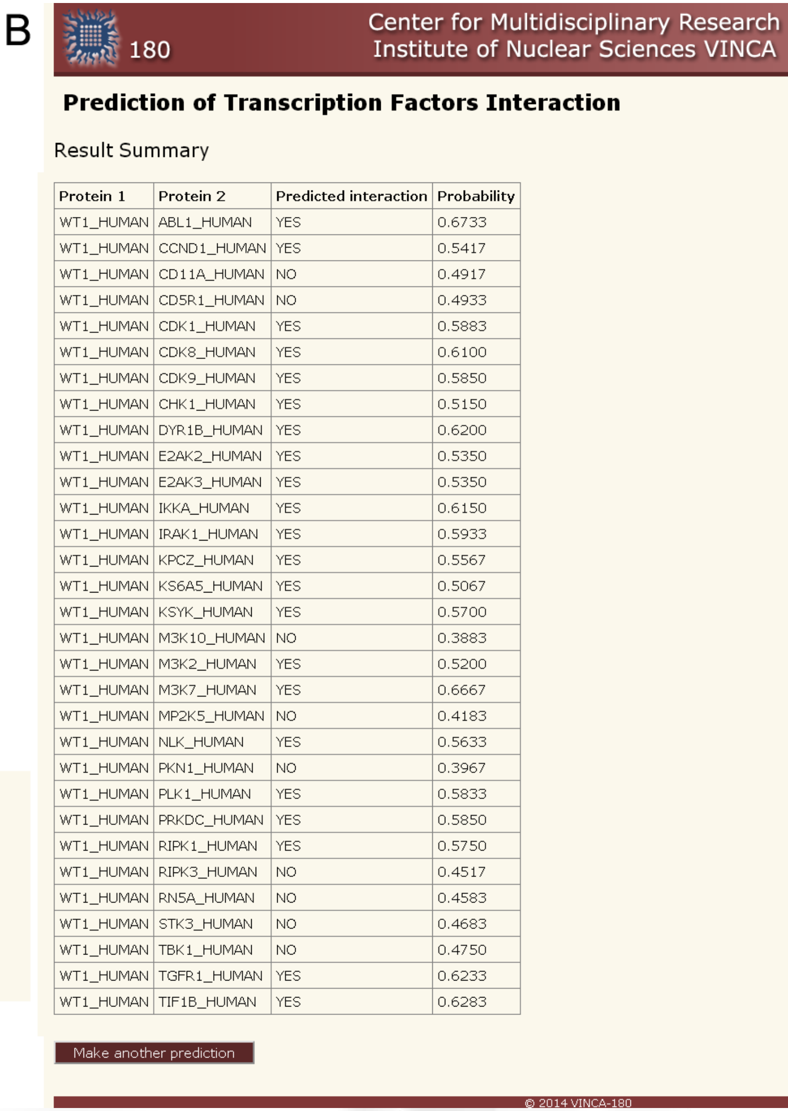
Examples

You can see the valid input formats in the next few examples. Click on the button and example data will fill the boxes in the form above.

- Example 1 User-input pairs: WT1 interactions with protein kinases in transcriptional regulation.
- Example 2 All possible pair combinations: Interactions in breast cancer transcriptional regulation. Result table will be sorted by the prediction value.

Clear Clear all fields.

© 2014 VINCA-180


B
180
Center for Multidisciplinary Research
Institute of Nuclear Sciences VINCA

Prediction of Transcription Factors Interaction

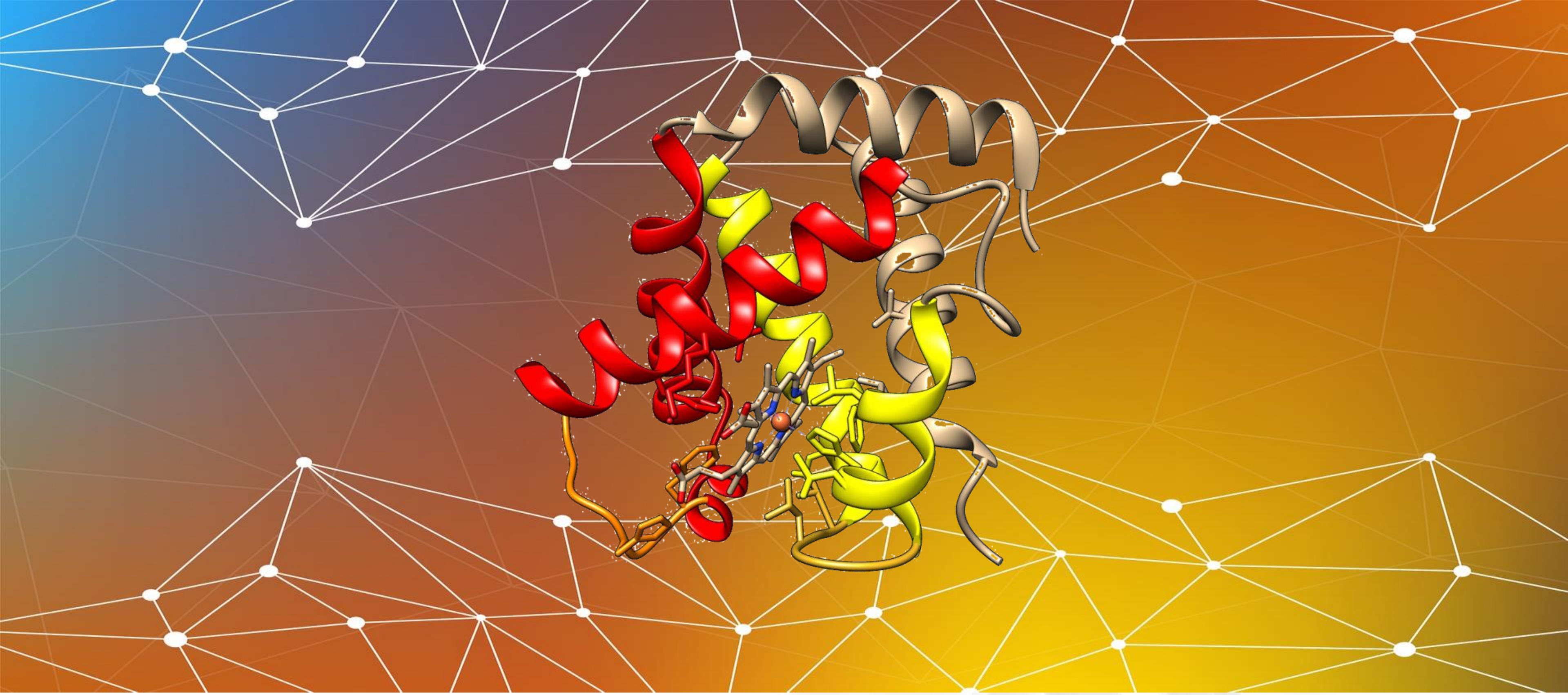
Result Summary

Protein 1	Protein 2	Predicted interaction	Probability
WT1_HUMAN	ABL1_HUMAN	YES	0.6733
WT1_HUMAN	CCND1_HUMAN	YES	0.5417
WT1_HUMAN	CD11A_HUMAN	NO	0.4917
WT1_HUMAN	CD5R1_HUMAN	NO	0.4933
WT1_HUMAN	CDK1_HUMAN	YES	0.5883
WT1_HUMAN	CDK8_HUMAN	YES	0.6100
WT1_HUMAN	CDK9_HUMAN	YES	0.5850
WT1_HUMAN	CHK1_HUMAN	YES	0.5150
WT1_HUMAN	DYR1B_HUMAN	YES	0.6200
WT1_HUMAN	E2AK2_HUMAN	YES	0.5350
WT1_HUMAN	E2AK3_HUMAN	YES	0.5350
WT1_HUMAN	IKKA_HUMAN	YES	0.6150
WT1_HUMAN	IRAK1_HUMAN	YES	0.5933
WT1_HUMAN	KPCZ_HUMAN	YES	0.5567
WT1_HUMAN	KS6A5_HUMAN	YES	0.5067
WT1_HUMAN	KSYK_HUMAN	YES	0.5700
WT1_HUMAN	M3K10_HUMAN	NO	0.3883
WT1_HUMAN	M3K2_HUMAN	YES	0.5200
WT1_HUMAN	M3K7_HUMAN	YES	0.6667
WT1_HUMAN	MP2K5_HUMAN	NO	0.4183
WT1_HUMAN	NLK_HUMAN	YES	0.5633
WT1_HUMAN	PKN1_HUMAN	NO	0.3967
WT1_HUMAN	PLK1_HUMAN	YES	0.5833
WT1_HUMAN	PRKDC_HUMAN	YES	0.5850
WT1_HUMAN	RIPK1_HUMAN	YES	0.5750
WT1_HUMAN	RIPK3_HUMAN	NO	0.4517
WT1_HUMAN	RN5A_HUMAN	NO	0.4583
WT1_HUMAN	STK3_HUMAN	NO	0.4683
WT1_HUMAN	TBK1_HUMAN	NO	0.4750
WT1_HUMAN	TGFR1_HUMAN	YES	0.6233
WT1_HUMAN	TIF1B_HUMAN	YES	0.6283

© 2014 VINCA-180

TRI_tool Web Interface (A) Front page of TRI_tool web application where users can input the protein sequences in a FASTA format and to choose either automatic combination in pairs or to add protein pairs of interest to the input information. (B) TRI_tool results page.

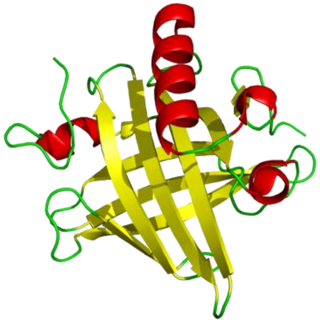
Perovic et al., Bioinformatics, 2017



Protein function prediction problem

Ontological annotation of proteins

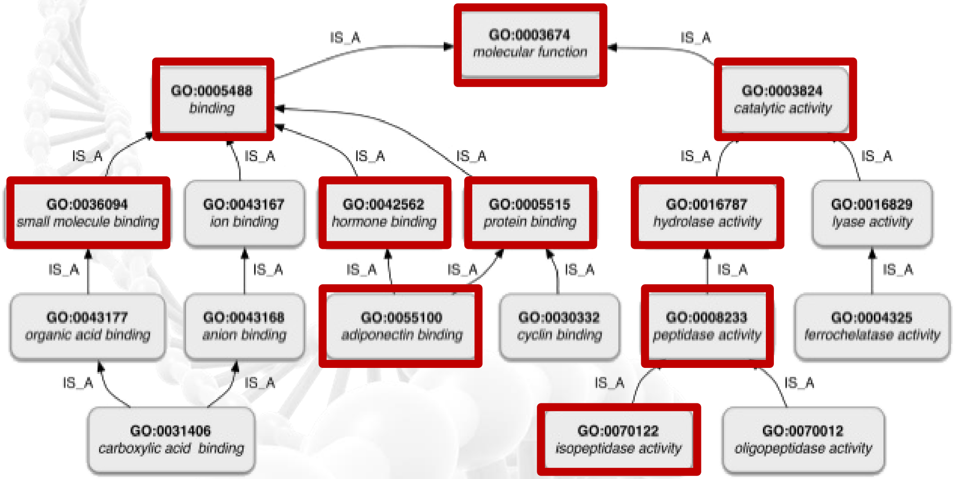
Protein



Assign/predict subgraph

Direct acyclic graph (DAG) of annotations

Example from Molecular Function ontology



Multi-label classification problem

Challenges

- Inconsistent experiments – in vitro, in vivo
- Biased and incomplete biological data

Why this matters

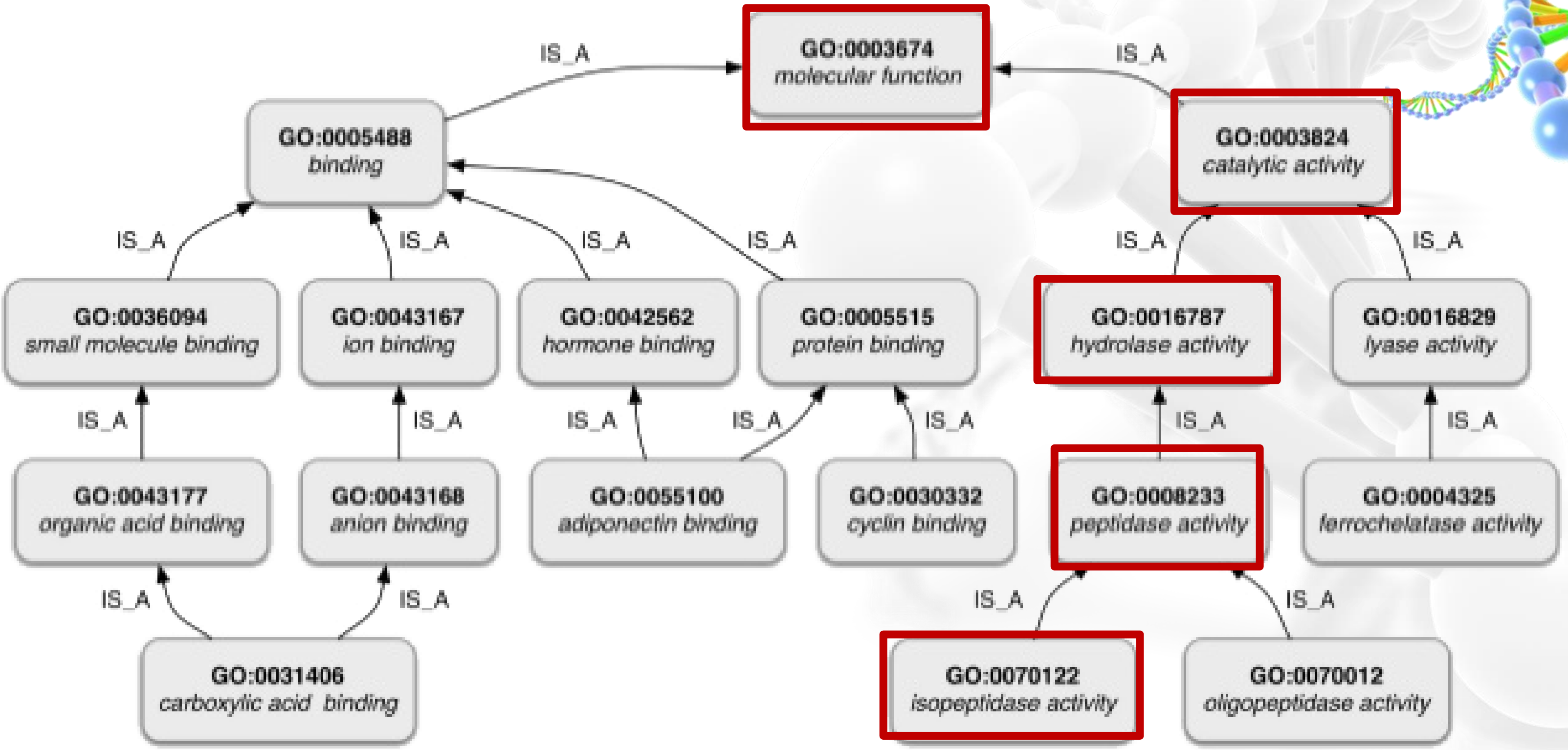
- Understand molecular mechanisms and cellular processes
- Mutation assessment, drug design…

Gene Ontologies (GO)

Gene Ontology (GO) is a term that describes gene product in three domains (across all spieces):

1. **Molecular function** - molecular activities of gene products
2. **Cellular component** - where gene products are active
3. **Biological process** - pathways and larger processes made up of the activities of multiple gene products.

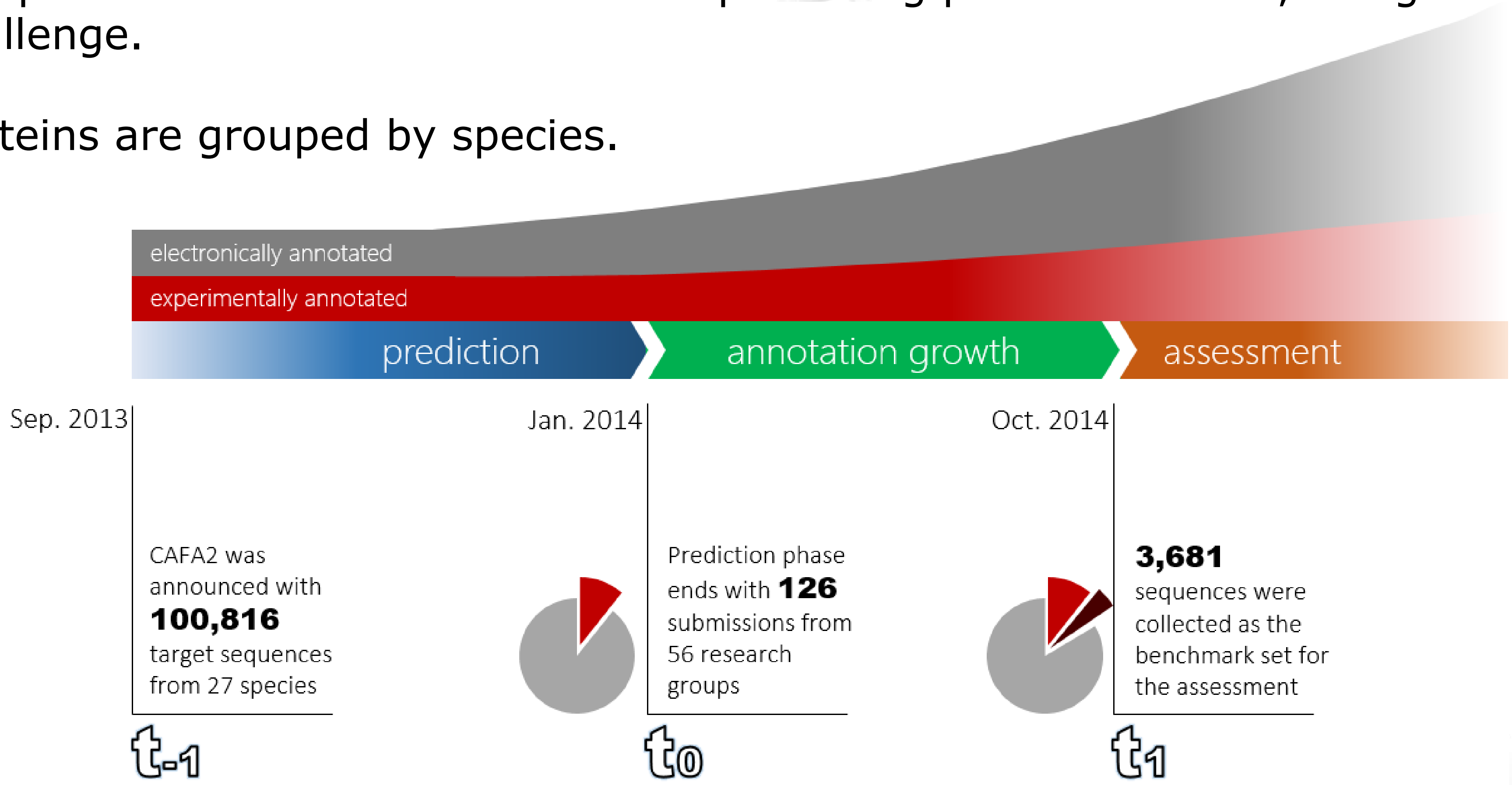
Vocabulary of GOs is structured in a **graph**



The CAFA Challenge

Critical Assessment of protein Function Annotation algorithms (CAFA) is an experiment designed to provide a large-scale assessment of computational methods dedicated to predicting protein function, using a time challenge.

Proteins are grouped by species.



CAFA2 was announced with **100,816** target sequences from 27 species

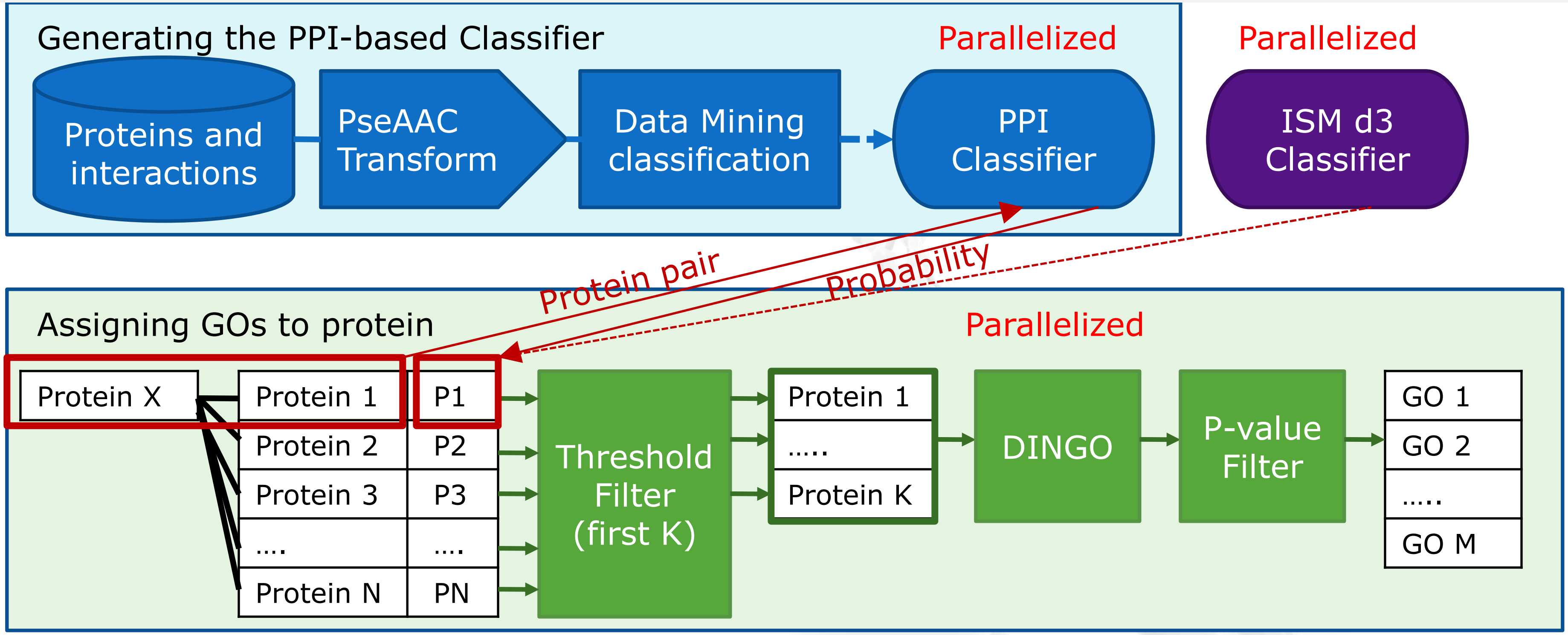
Prediction phase ends with **126** submissions from 56 research groups

3,681 sequences were collected as the benchmark set for the assessment

*Jiang Y., Oron T., Clarck W.T. et al. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. Genome Biol. 2016;17(1):184. (**IF=13.2**)*

The CAFA Challenge - Prediction model

Algorithm



Davidovic R, Perovic V, Gemovic B and Veljkovic N. (2019) ***DiNGO****: standalone application for Gene Ontology and Human Phenotype Ontology term enrichment analysis. Bioinformatics. In submission.*

DiNGO software page: https://www.vin.bg.ac.rs/180/tools/DiNGO.php

Big Data in 'Assigning GOs' step

20 species, total ~550K proteins

Human organism: 20K proteins → 400M pairs: PPI based model → (x140) 56B numbers ~ 0.45TB

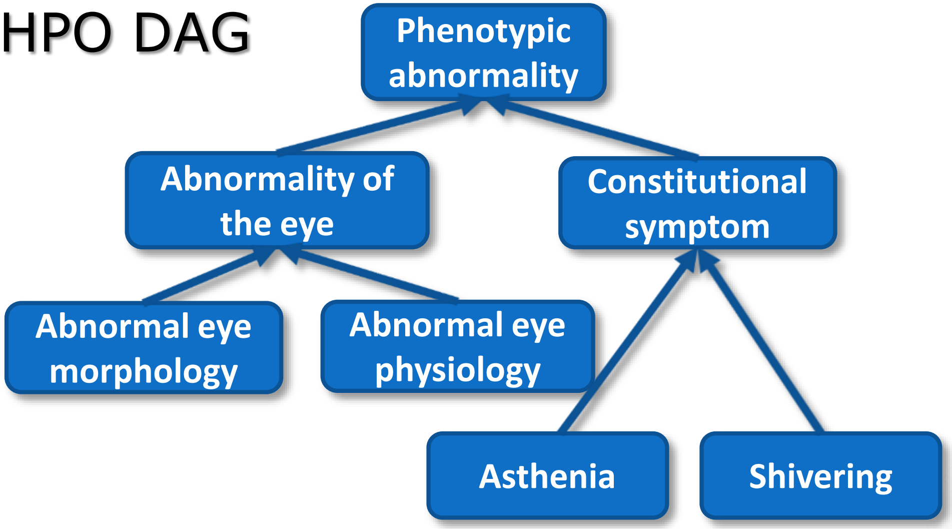
ISM d3 based → (x8000) 3.2T numbers ~ 25TB

*Zhou N., Jiang Y., Nguyen H., Hamid M. et al. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. Genome Biol. 2019; Accepted. (****IF=13.2****)*

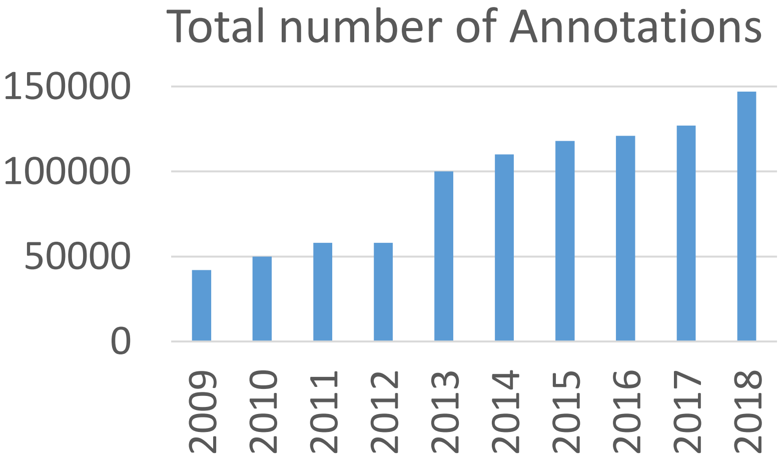
The Human Phenotype Ontology (HPO)

Database of phenotypic abnormalities in human diseases

HPO DAG



- Difficult to analyze a patient information by computerized approaches.
- Phenotypic information - unstructured clinical notes (traditionally)
- HPO standardizes clinical feature descriptions, in a way that is consistent and computer-readable



HPO Mar-2018

Subontology	Terms	Proteins
Phenotypic abnormality	6953	3645
Mode of Inheritance	21	3333
Clinical modifier	22	1263
Aging/Mortality	6	226

Not many tools for HPO annotation prediction

PHENOstruct – M1

- Based on structured support vector machine (SSVM)
- Features:
  - Network data (PPI, co-expression, co-occurrence, etc.) from BioGRID, STRING and GeneMANIA
  - Gene Ontology (GO)
  - Literature
  - Disease variants (UniProt)

Kahanda et al., F1000Research, 2015

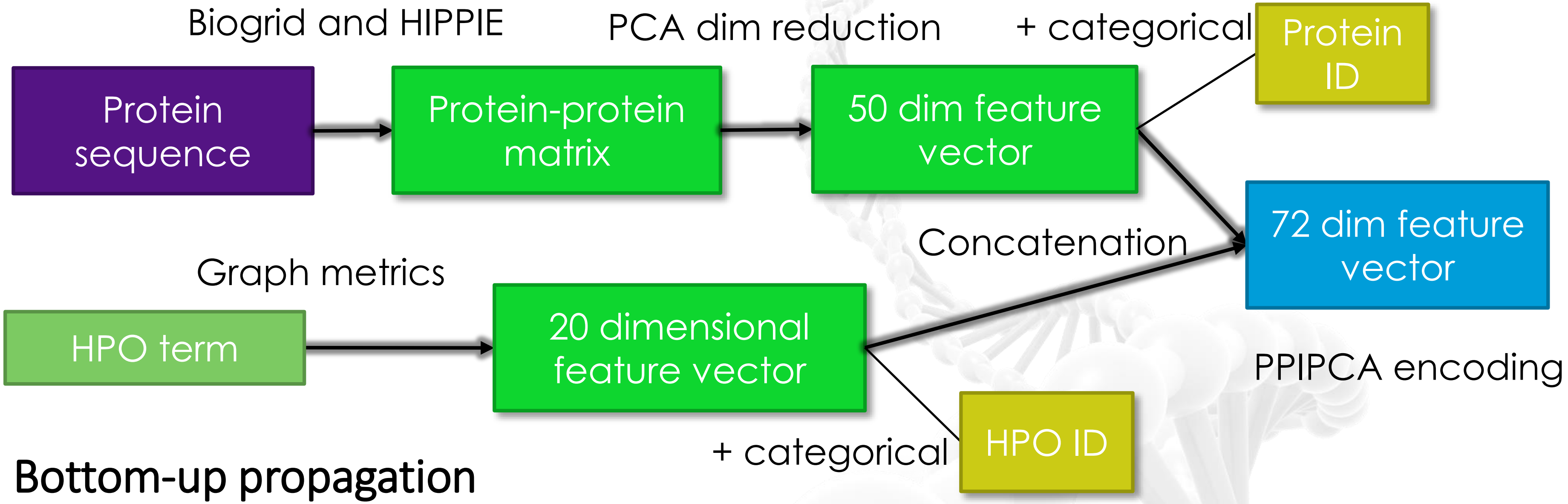
HEMDAG – M2

- Hierarchical top down (HTD) and True path rule (TPR) propagation algorithms
- SVM and RANKS ML methods
- Features:
  - Network data (PPI, co-expression, co-occurrence, etc.) from BioGRID and STRING
  - Gene Ontology (GO)
  - OMIM annotations

Notaro et al., BMC Bioinformatics, 2017

HPO prediction
Proteome-wide approach

MuFEnsHPO model for HPO prediction



Biogrid and HIPPIE

PCA dim reduction

+ categorical

Protein sequence

Protein-protein matrix

50 dim feature vector

Protein ID

Concatenation

72 dim feature vector

Graph metrics

HPO term

20 dimensional feature vector

PPIPCA encoding

+ categorical

HPO ID

Bottom-up propagation

Binary classifier

Negative examples = annotations complement

Ensemble model

- Random forest
- Gradient boosted machine
- Generalized linear model

Evaluation

5-fold CV protein centric

Dataset size

Phenotypic abnormality: ~**25M** ex

Mode of Inheritance: ~**28K** ex

Clinical modifier: **70K** ex

Aging/Mortality: **1.4K** ex

Performances of *GraPPI* model

Mode of Inheritance (v2014)

Method	max F	Precision	Recall
M1	0.74	0.68	0.81
M2	0.69	0.59	**0.82**
MuFEnsHPO	**0.75**	**0.69**	**0.82**

Clinical modifier (v2014)

Method	max F	Precision	Recall
M1	0.39	0.31	0.52
M2	0.48	0.38	**0.66**
MuFEnsHPO	**0.52**	**0.48**	0.56

Phenotypic abnormality (v2014)

Method	max F	Precision	Recall
M1	0.42	0.35	**0.56**
M2	**0.44**	**0.38**	0.51
MuFEnsHPO	0.37	0.34	0.40

Aging/Mortality (v2018)

Method	max F	Precision	Recall
MuFEnsHPO	**0.61**	**0.57**	**0.62**

Evaluation of predictions on HPO updated release

Data sets

Dataset	Term-protein pairs	Terms	Proteins
Train HPO jan-2014	6,841,110	2,445	2,797
Test apr-2016	1,484,115	2,445	608

- all annotations -

Notaro et al. Prediction of Human Phenotype Ontology terms by means of hierarchical ensemble methods. BMC Bioinformatics (2017) 18:449

Performance

Method	max F	Precision	Recall	Training time
M1	0.3635	0.3040	**0.4519**	18 hours
M2	0.3826	0.3512	0.4202	3 hours
MuFEnsHPO	0.3775	0.3484	0.4119	**21 min**
M2 + MuFEnsHPO	**0.3946**	**0.3530**	0.4474	

SUMMARY

Summary

Sequence is universal and reliable protein representation, suitable for automatic predictions

Protein-protein interaction (PPI) prediction

↗.. Improved performance with amino acid **physico-chemical characteristics**

↗...... with **protein profile** data

↗........ with **graph features**

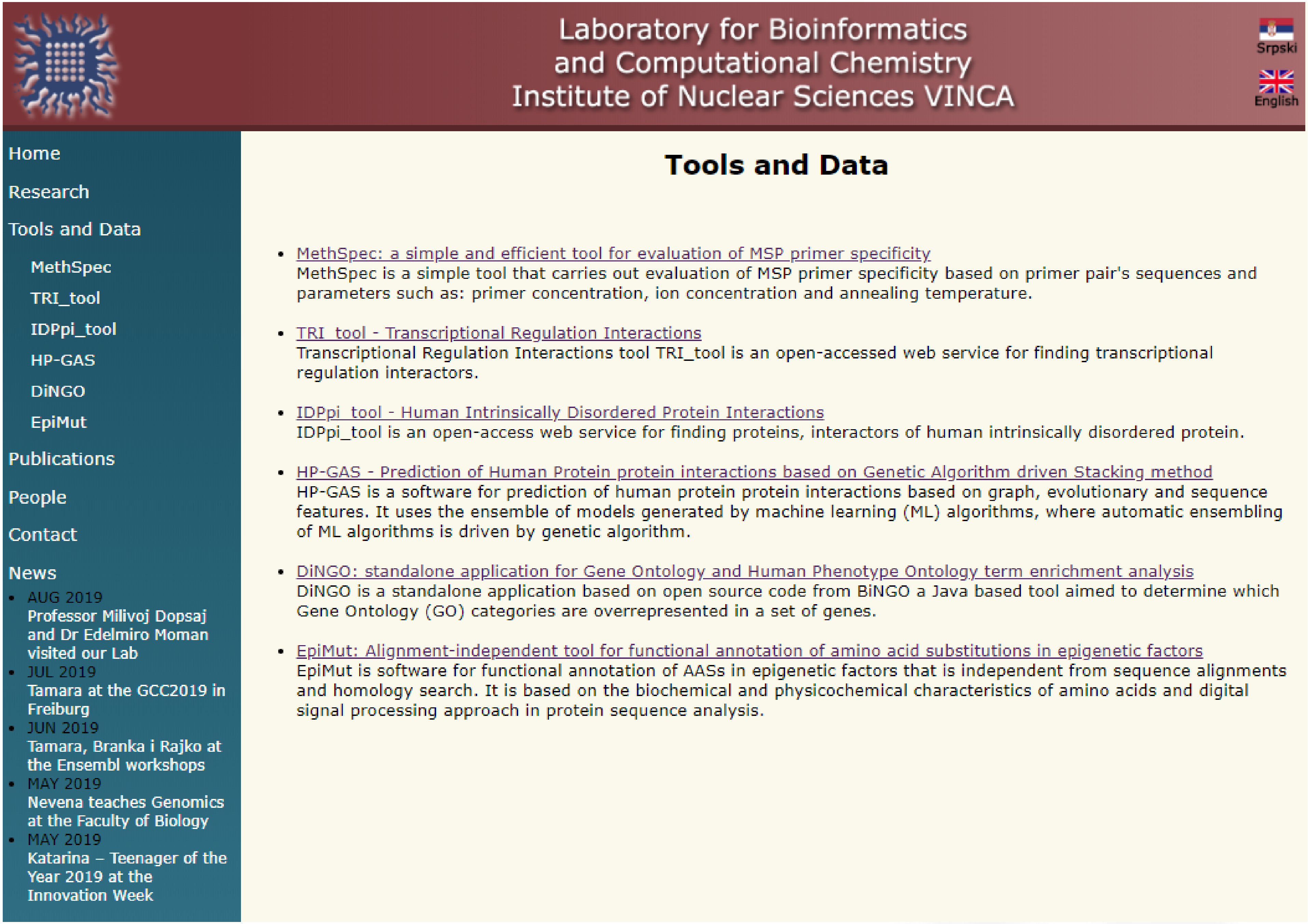
Multi feature **ensemble** of different ML algorithms significantly improved the PPI predictive performances

Human Phenotype Ontology (HPO) prediction models based on sequence, Graph metrics and PPI data have **satisfactory** predictive performance

All MuFEns methods are **time efficient**

IDPs, are currently largely missing from HPO, but since they are involved in many disease, they will be in the future more present and curated in HPO

www.vinca.rs/180



Laboratory for Bioinformatics
and Computational Chemistry
Institute of Nuclear Sciences VINCA

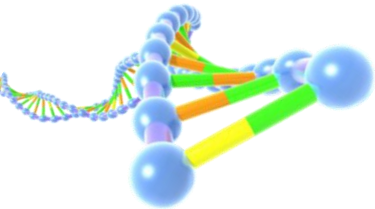
Tools and Data

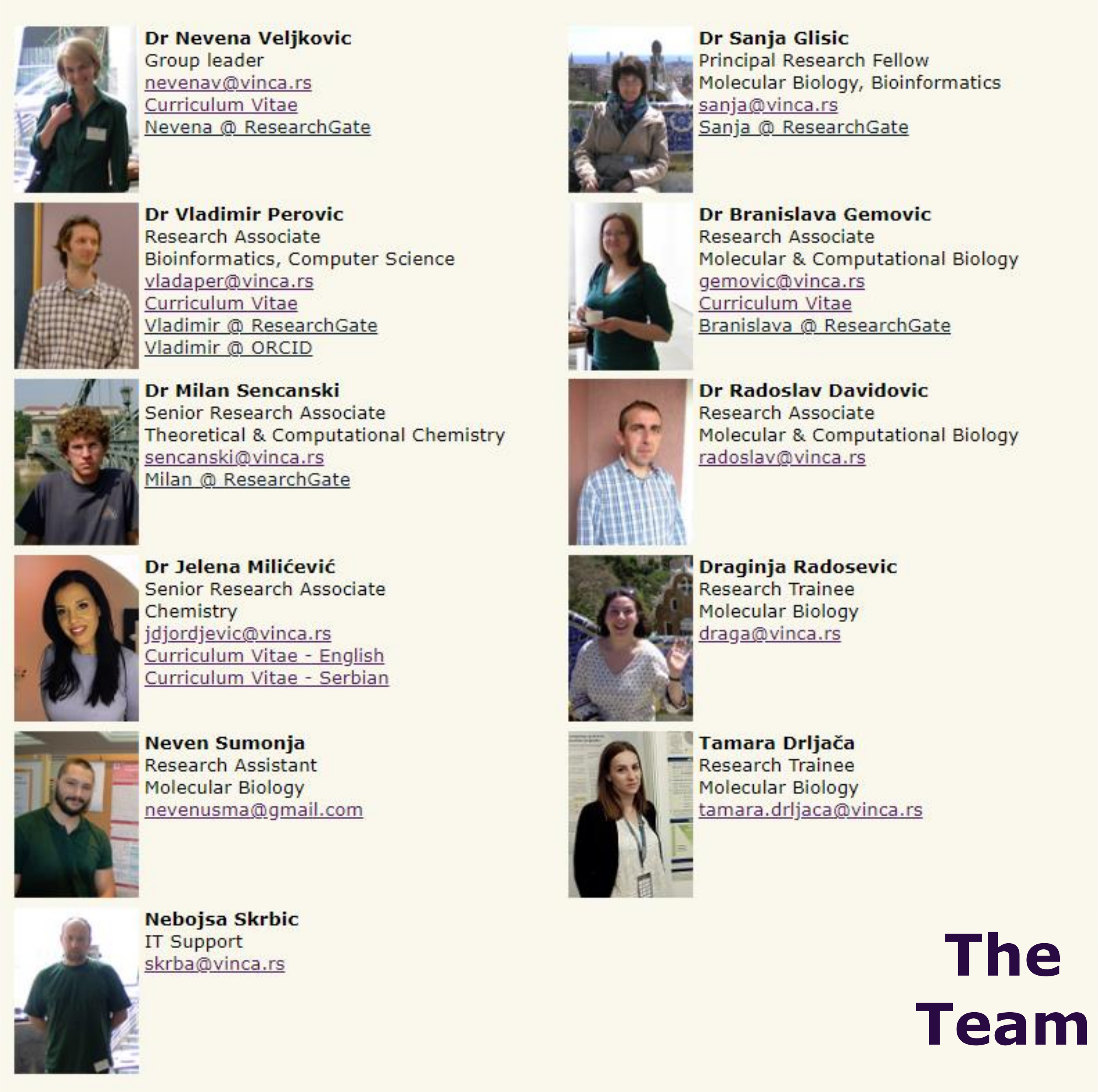
- MethSpec: a simple and efficient tool for evaluation of MSP primer specificity
  MethSpec is a simple tool that carries out evaluation of MSP primer specificity based on primer pair's sequences and parameters such as: primer concentration, ion concentration and annealing temperature.
- TRI_tool - Transcriptional Regulation Interactions
  Transcriptional Regulation Interactions tool TRI_tool is an open-accessed web service for finding transcriptional regulation interactors.
- IDPpi_tool - Human Intrinsically Disordered Protein Interactions
  IDPpi_tool is an open-access web service for finding proteins, interactors of human intrinsically disordered protein.
- HP-GAS - Prediction of Human Protein protein interactions based on Genetic Algorithm driven Stacking method
  HP-GAS is a software for prediction of human protein protein interactions based on graph, evolutionary and sequence features. It uses the ensemble of models generated by machine learning (ML) algorithms, where automatic ensembling of ML algorithms is driven by genetic algorithm.
- DiNGO: standalone application for Gene Ontology and Human Phenotype Ontology term enrichment analysis
  DiNGO is a standalone application based on open source code from BiNGO a Java based tool aimed to determine which Gene Ontology (GO) categories are overrepresented in a set of genes.
- EpiMut: Alignment-independent tool for functional annotation of amino acid substitutions in epigenetic factors
  EpiMut is software for functional annotation of AASs in epigenetic factors that is independent from sequence alignments and homology search. It is based on the biochemical and physicochemical characteristics of amino acids and digital signal processing approach in protein sequence analysis.

News

- AUG 2019
  Professor Milivoj Dopsaj and Dr Edelmiro Moman visited our Lab
- JUL 2019
  Tamara at the GCC2019 in Freiburg
- JUN 2019
  Tamara, Branka i Rajko at the Ensembl workshops
- MAY 2019
  Nevena teaches Genomics at the Faculty of Biology
- MAY 2019
  Katarina – Teenager of the Year 2019 at the Innovation Week

The Team





Dr Nevena Veljkovic
Group leader
nevenav@vinca.rs
Curriculum Vitae
Nevena @ ResearchGate

Dr Sanja Glisic
Principal Research Fellow
Molecular Biology, Bioinformatics
sanja@vinca.rs
Sanja @ ResearchGate

Dr Vladimir Perovic
Research Associate
Bioinformatics, Computer Science
vladaper@vinca.rs
Curriculum Vitae
Vladimir @ ResearchGate
Vladimir @ ORCID

Dr Branislava Gemovic
Research Associate
Molecular & Computational Biology
gemovic@vinca.rs
Curriculum Vitae
Branislava @ ResearchGate

Dr Milan Sencanski
Senior Research Associate
Theoretical & Computational Chemistry
sencanski@vinca.rs
Milan @ ResearchGate

Dr Radoslav Davidovic
Research Associate
Molecular & Computational Biology
radoslav@vinca.rs

Dr Jelena Milićević
Senior Research Associate
Chemistry
jdjordjevic@vinca.rs
Curriculum Vitae - English
Curriculum Vitae - Serbian

Draginja Radosevic
Research Trainee
Molecular Biology
draga@vinca.rs

Neven Sumonja
Research Assistant
Molecular Biology
nevenusma@gmail.com

Tamara Drljača
Research Trainee
Molecular Biology
tamara.drljaca@vinca.rs

Nebojsa Skrbic
IT Support
skrba@vinca.rs

Acknowledgements

Institute of Nuclear Sciences Vinča

Nevena Veljkovic

Neven Sumonja

Branislava Gemovic

Radoslav Davidovic

Sanja Glisic

Veljko Veljkovic

Ministry of
Education, Science and
Technological Development of the
Republic of Serbia

МИНИСТАРСТВО ПРОСВЕТЕ,
НАУКЕ И ТЕХНОЛОШКОГ РАЗВОЈА

cost
EUROPEAN COOPERATION
IN SCIENCE AND TECHNOLOGY

University of
BRISTOL

Lindsey Marsh

Stefan Roberts

МАТЕМАТИЧКИ ФАКУЛТЕТ
У БЕОГРАДУ
1873

Mladen Nikolic

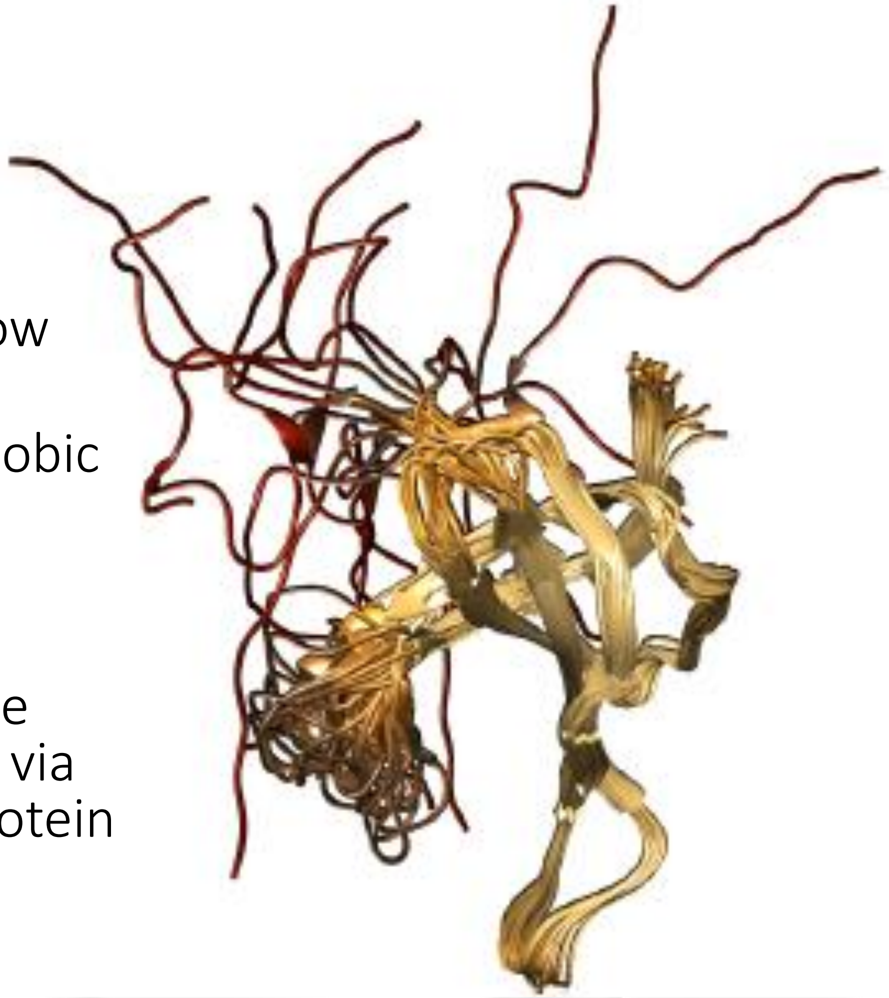
Jovana Kovacevic

THANK YOU

APPENDIX

Intrinsically disordered proteins (IDPs)

- The lack of a fixed tertiary structure
- ~33% IDPs biologically functional in Eukaryota
- Biased amino acid composition and low sequence complexity
  - low proportions of bulky hydrophobic amino acids
  - high proportions of charged and hydrophilic amino acids
- Functionally important: involved in the regulation of key biological processes via binding to significantly augmented protein partners.

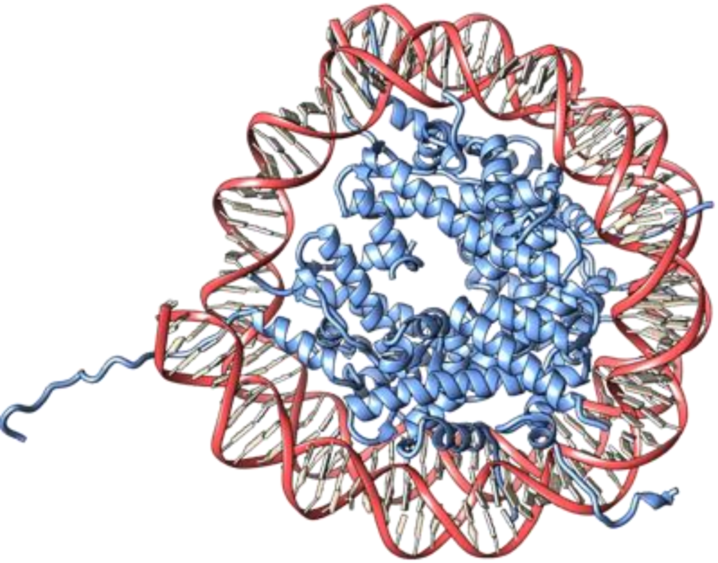


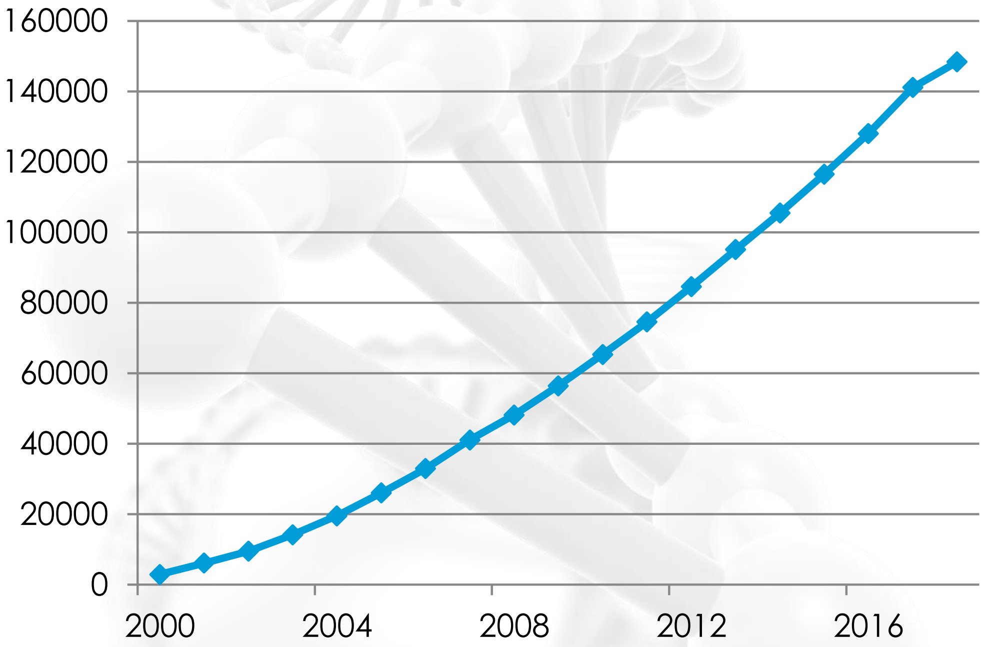
Protein Structures Database

wwPDB – worldwide Protein Data Bank *https://www.wwpdb.org*

- The single repository of information about the 3D structures of proteins, nucleic acids, and complex assemblies
- Established in 1971 in Uptown, New York, US

148,626 structures



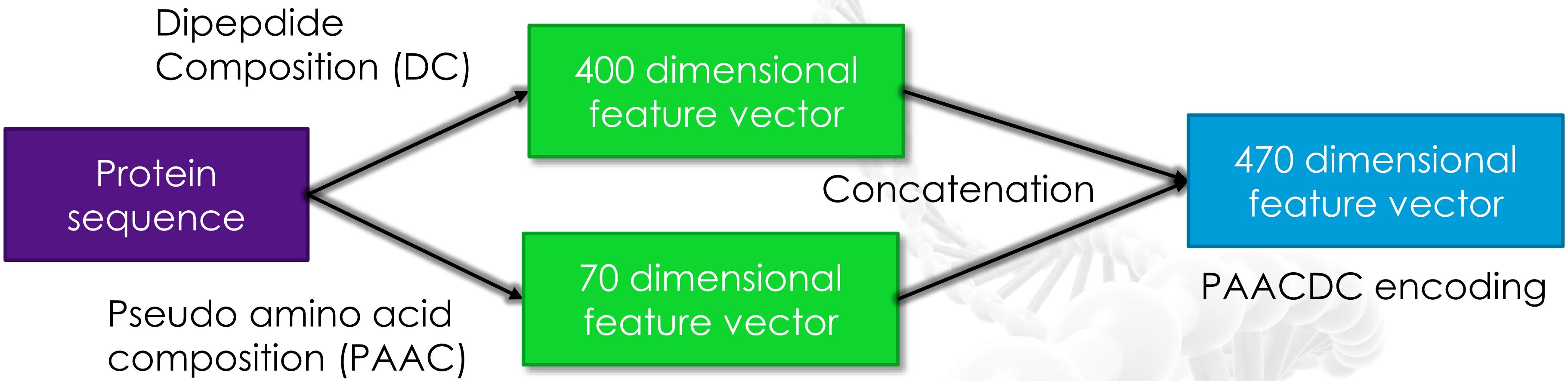


Statistics for PDB structures that are deposited and processed by year

HPO prediction
Class-specific approach

HPO prediction for Intrinsically Disorder Proteins

IDPs representation – PAACDC features



Dipepdide Composition (DC)

Protein sequence

400 dimensional feature vector

70 dimensional feature vector

Pseudo amino acid composition (PAAC)

Concatenation

470 dimensional feature vector

PAACDC encoding

PAAC is using five disorder characteristic propensity scales:

- **TOP-IDP** scale (ranks residues by the their propensity to endorse order or disorder)
- **B-values** (flexibility parameters for each residue surrounded by two inflexible neighbours)
- **FoldUnfold** scale (capacity of amino acid residues to form a sufficient number of contacts in a globular state)
- **DisProt** scale (statistical difference in the residue compositions of ordered proteins and IDPs)
- **Net charge** scale

Performance of annotation predictions on IDPs

PHENOstruct with PAACDC features

Clinical modifier

Method	max F	Precision	Recall
M1	0.4776	0.3429	**0.7866**
M1+ PAACDC	**0.5220**	**0.4503**	0.6208

Mode of Inheritance

Method	max F	Precision	Recall
M1	0.7682	0.6939	**0.8605**
M1+ PAACDC	**0.7750**	**0.7648**	0.7852

Performance of PAACDC model

Clinical modifier

Method	max F	Precision	Recall
M1	0.4776	0.3429	**0.7866**
PAACDC	**0.5729**	**0.6750**	0.4975

Mode of Inheritance

Method	max F	Precision	Recall
M1	**0.7682**	**0.6939**	**0.8605**
PAACDC	0.7122	0.6370	0.8075