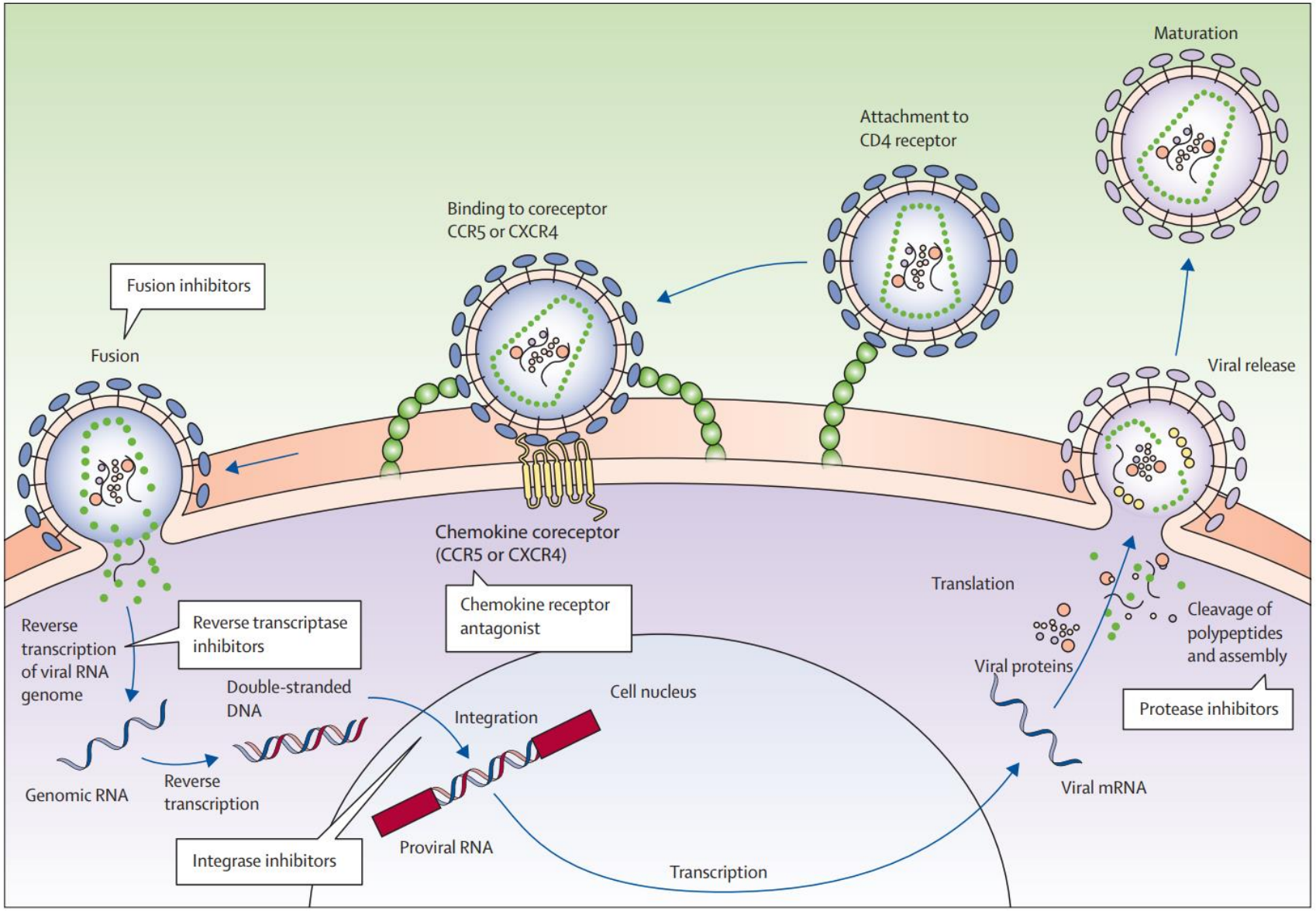
XXVIII Symposium on Bioinformatics and Computer-Aided Drug Discovery

INFORMATION EXTRACTION FROM TEXTS: ANTIVIRAL AGENTS ACTIVE AGAINST VIRUS OR HOST PROTEINS

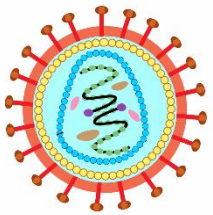
N.Y. Biziukova, O.A. Tarasova, D.A. Filimonov, V.V. Poroikov

Institute of Biomedical Chemistry, Moscow, Russia

ИБМХ



Antiretroviral therapy

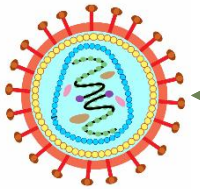


HIV key proteins

Human organism

Host dependency factors

Host restriction factors



HIV

Virus-host interaction mechanisms

Pathogenesis of HIV-associated diseases

Human genes polymorphisms associated with susceptibility/resistance to HIV-infection

Other HIV-related topics

> 300 000 results!

Purpose

Identification of chemical compounds that act on human and virus proteins involved in the HIV infection pathogenesis mechanisms

Tasks

- Development of chemical and protein named entity recognition algorithm
- Identification of proteins and genes involved HIV-host interactions based on the developed algorithm
- Search for chemical compounds that are known to affect key proteins in HIV infection pathogenesis
- Representation of work results as graphical interaction networks

Chemical and protein named entity recognition algorithm

Annotated text corpora

CHEMDNER
- 10 000 abstracts
- Chemical named entity annotations

DrugProt
- 15 000 abstracts
- Chemical and proteins (genes) named entity annotations

Tokenization

Chemical NER
- Punctuation marks and word separators

Proteins and genes NER
- Punctuation marks

Token labeling

Chemical NER
- SOBIE

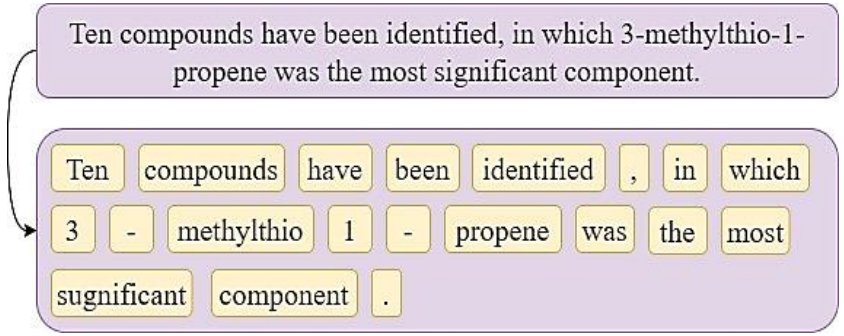
Proteins and genes NER
- SOBIE

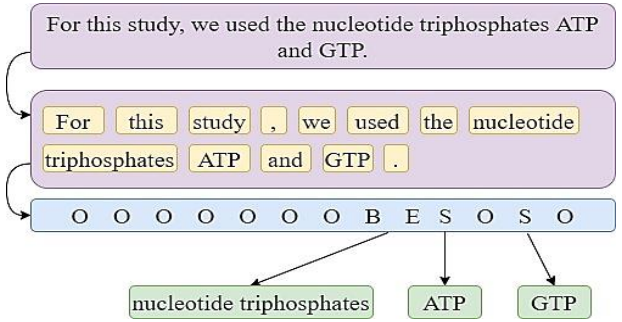
CHEMDNER

```
22075688   A   64   74   (25)Mg(2+)      FORMULA
22075688   A   736  739  ATP             ABBREVIATION
22546614   A   780  787  ethanol         SYSTEMATIC
22566187   A   639  650  polyphenols     FAMILY
```

DrugProt

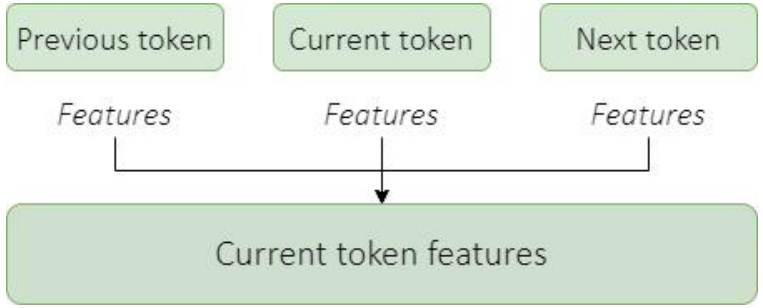
```
23538162   T10   CHEMICAL   947   957   ICI 82,780
23538162   T11   CHEMICAL   1002  1005  Rg1
23538162   T12   CHEMICAL   72    87    ginsenoside Rg1
23538162   T13   GENE-Y     1330  1337  Aβ25-35
```





S – Single
B – Begin
I – Inside
E – End
O - Out

Feature	Type	Meaning
word	string	Token
lower	string	Token (lower case)
isUpper	Boolean	Is token in the upper case
isTitle	Boolean	Is token a title (first character in the upper case)
isDigit	Boolean	Is token a digit
hasDigits	Boolean	Does token have digits
isNonSpecific	Boolean	Is token a non-specific term
isStopWord	Boolean	Is token a stop-word
hasSymbols	Boolean	Does token have symbols
word[n-3:n]	string	Last three characters of the token
word[n-2:n]	string	Last to characters of the token
firstChar	string	Token first character
length	integer	Number of characters in the token
posTag	string	Token part of speech tag



Algorithm: Conditional Random Fields (CRF)

Realization: Python 3.10

- NLTK
- Sklearn_crfcuite

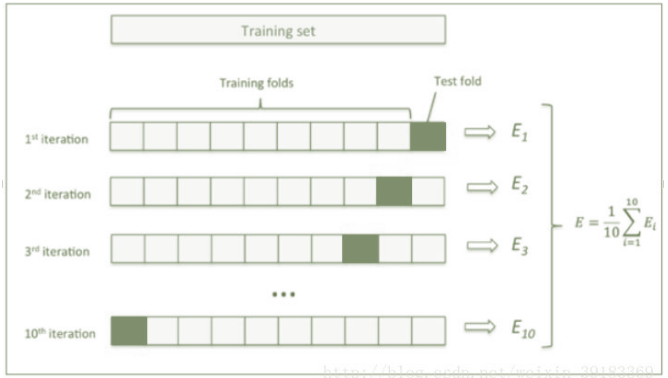
Recognition accuracy

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F1\text{-}score = \frac{2 * precision * recall}{precision + recall}$$

- Five-fold cross validation



- Manual analysis of recognition results on a texts sample (100 abstracts)

Chemical named entity recognition

Five-fold cross-validation

	Precision	Recall	F1-score
S	0,91	0,84	0,87
O	0,99	0,99	0,99
B	0,87	0,80	0,83
I	0,92	0,89	0,91
E	0,88	0,81	0,85
Avg	**0,91**	**0,87**	**0,89**

Manual annotation (Test set)

	Precision	Recall	F1-score
CNE	0,83	0,94	0,88

Protein (gene) named entity recognition

Five-fold cross-validation

	Precision	Recall	F1-score
S	0,86	0,83	0,84
O	0,97	0,98	0,98
B	0,83	0,78	0,80
I	0,84	0,81	0,83
E	0,83	0,78	0,81
Avg	**0,87**	**0,84**	**0,85**

Manual annotation (Test set)

	Precision	Recall	F1-score
PNE	0,84	0,79	0,81

Identification of proteins and genes involved HIV-host interactions

"Host-Pathogen Interactions"[Mesh] *AND* Subheadings

"HIV"[Mesh]

1960 abstracts

Sentences

Protein and genes named entity recognition

Sentences with one protein named entity

Sentence	Phrase	Pattern	Direction
Recent findings suggest that HIV-1 viral protein R (Vpr) **interacts with** some of the host innate antiviral factors, such as heat shock proteins, and plays an active role as a viral pathogenic factor.	**interacts with**	1 phrase 2	No matter
However, Vpr overcomes these heat-stress-like responses **by preventing** heat shock factor-1 (HSF-1)-mediated activation of heat shock proteins.	by preventing	1 phrase 2	1 --> 2
In addition to heat stress response proteins, we will discuss **interactions of** Vpr with other proteins, such as EF2 and Skp1/GSK3, their involvements in cellular responses to Vpr, as well as strategies to develop novel antiviral therapies aimed at enhancing anti-Vpr responses of the host cell.	interactions of	phrase 1 2	No matter
Differential **regulation of** indoleamine-2,3-dioxygenase (IDO) by HIV type 1 clade B and C Tat protein.	regulation of	phrase 1 2	2 --> 1
We hypothesize that HIV-1 clade B and C Tat proteins might exert differential effects on human primary astrocytes **by the upregulation of** the IDO gene and protein expression as well as its activity and production of the neurotoxin KYN.	by the upregulation of	1 phrase 2	1 --> 2
Our results indicate that HIV-1 clade B Tat protein significantly **upregulated** the IDO gene and protein expression, IDO enzyme activity, as well as KYN concentration compared to HIV-1 clade C Tat protein.	upregulated	1 phrase 2	1 --> 2
Thus, our studies for the first time demonstrate that HIV-1 clade B Tat protein in human primary astrocytes appears to **increase** the level of neuropathogenic agents, such as IDO and KYN, as compared to HIV-1 clade C Tat protein.	increase	1 phrase 2	1 --> 2
HA **binds to** glycans-containing receptors with terminal sialic acid (SA).	binds to	1 phrase 2	No matter
Filamin **binds to** both CD4 and CXCR4 in a manner promoted by signaling of the HIV gp120 glycoprotein.	binds to	1 phrase 2	No matter
ERM proteins **attach** actin filaments to the membrane and may promote polymerization of actin.	attach	1 phrase 2	No matter

Common term	Phrase	Pattern			Direction		
		1 phrase 2	phrase 1 2	1 2 phrase	1 --> 2	2 --> 1	No matter
--//--	-	1	0	0	0	0	1
--//--	/	1	0	0	0	0	1
acitvate	activating	1	0	0	1	0	0
	activation of	0	1	0	0	1	0
	activates	1	0	0	1	0	0
	activated by	1	0	0	0	1	0
	activate	1	0	0	1	0	0
attach	attach	1	0	0	0	0	1
	attaches	1	0	0	0	0	1
	attachment	1	1	0	0	0	1
	attached	1	0	0	0	0	1
bind	bind	1	0	0	0	0	1
	bind to	1	0	0	0	0	1
	binds	1	0	0	0	0	1
	binds to	1	0	0	0	0	1
	binding of	0	1	0	0	0	1

1960 abstracts

↓

1475 unique sentences with 2 or more protein named entities

↓

~3000 unique interactions (within the abstract)

Interacting proteins

↓

UniProt

Organism

UniProt ID

HIV

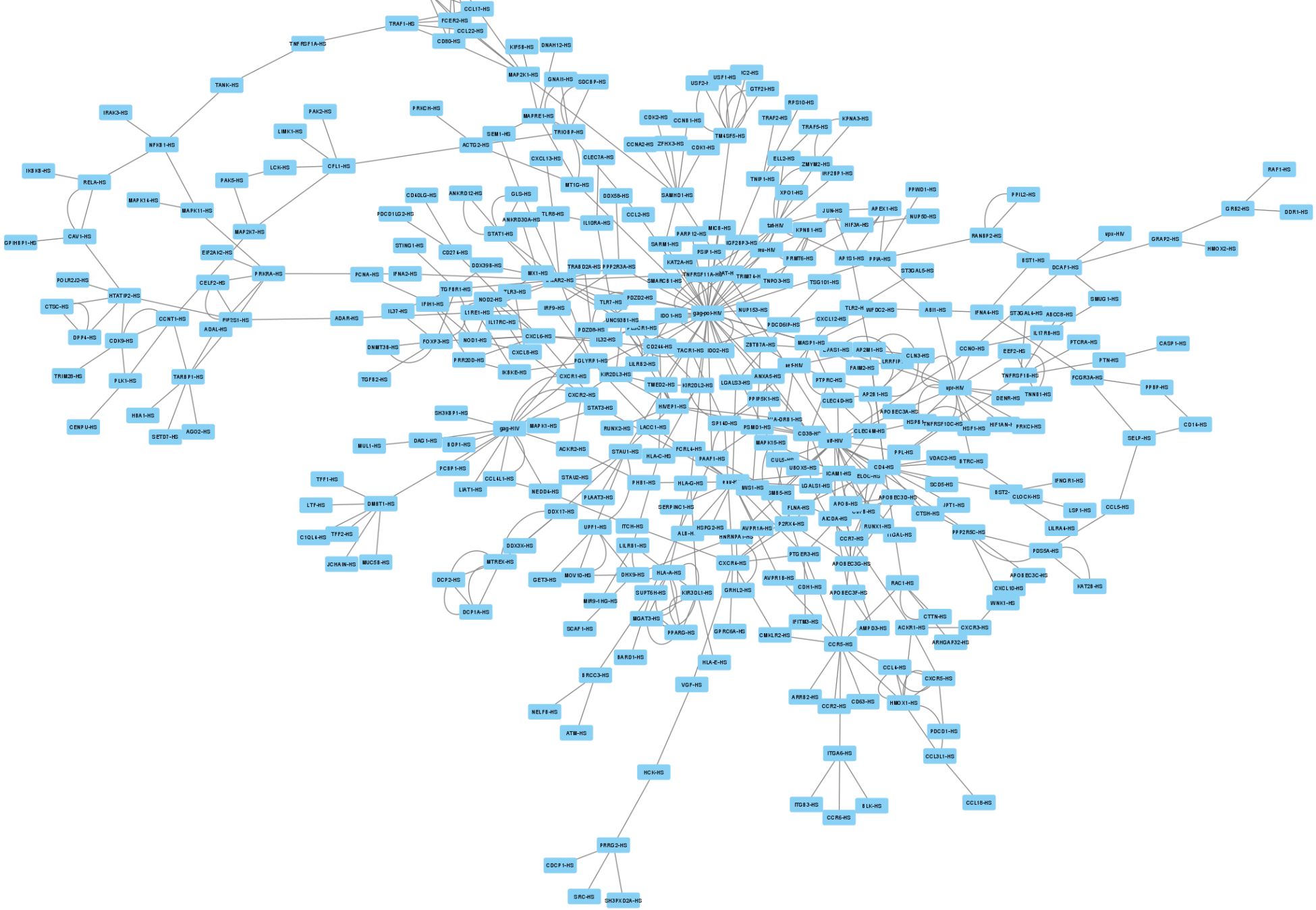
Human

Interacting pairs (TXT format) → Interacting pairs (SIF format)

APEX1-HS	pp	PPIA-HS	JUN-HS	AP1S1-HS		
SCD5-HS	pp	CD4-HS				
AICDA-HS	pp	APOBEC3G-HS	PSMB5-HS	vif-HIV	APOBEC3D-HS	APOBEC3F-HS
IL2RA-HS	pp	IL4R-HS	IL7R-HS	IL9R-HS	IL15RA-HS	IL21R-HS
IL4R-HS	pp	IL2RA-HS				
IL9R-HS	pp	IL2RA-HS				
IL15RA-HS	pp	IL2RA-HS				
IL21R-HS	pp	IL2RA-HS				
SCAF1-HS	pp	DHX9-HS				
PCBP1-HS	pp	gag-HIV				
PPIG-HS	pp					
gag-pol-HI\	pp	PPIA-HS	IFNAR2-HS	TRIM74-HS		
vpr-HIV	pp	vpr-HIV	HSPB1-HS			
HSPB1-HS	pp	vpr-HIV	vpr-HIV	HSF1-HS	vif-HIV	

Example of SIF format. HIV and HS labels were used to mark HIV and human proteins, respectively

Cytoscape



Part of HIV-host interactions network

The verification based on the prospective clinical study

HIV+ patients
(group of 10 individuals)

The longitudeness of HIV-infection

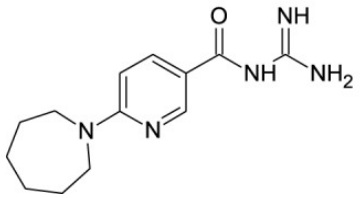
- Under 1 year
- More than 1 year

An analysis of differences in gene expression

Matching results of text analysis and clinical study

GENE	SYNONIMS	Gene Ontology	PMIDs	The associasion is known
CLEC5A	C-type lectin domain family 5 member A	Immune response	31867016	Нет
CXCL8	Interleukin-8, C-X-C motif chemokine 8, Emoctakin	Inflammatory response Regulation of gene expression	27227934 33610024	Да
FCGR2A	Low affinity immunoglobulin gamma Fc region receptor II-a, CD32, FcγRII	Immune response	25100508	Да
FPR1	fMet-Leu-Phe receptor, fMLP receptor	Inflammatory response	32093694	Да
TLR2	Toll-like receptor 2, CD282	Immune response Inflammatory response Regulation of gene expression	32093694 32516401 28730622	Да
NT5E	5'-nucleotidase	Inflammatory response	-	Нет
CD14	Monocyte differentiation antigen CD14	Immune response Inflammatory response	34211989 33487130	Да
CD86	T-lymphocyte activation antigen CD86	Immune response Negative regulation of T cell proliferation	34630420	Да
NAMPT	Nicotinamide phosphoribosyltransferase	Autophagy	-	Нет

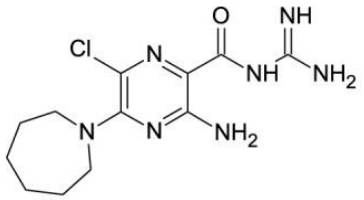
Antiviral compounds (HIV) that potentially act on the host proteins



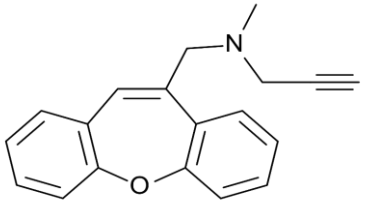
SM111

Micromolar inhibitors of HIV-1 replication; Possibly mediates downregulation of HIV-1's entry receptor CD4 and reduce expression of tetherin.

Philip Mwimanzi, J. Virol., 2016



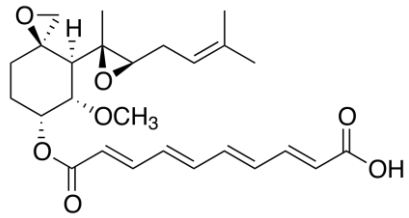
HMA



Omigapil

Modulation of GAPDH by Omigapil leads to dose-dependent inhibition of HIV, Dengue and Zika virus

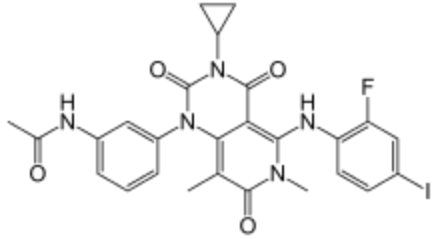
Trevor V. Gale, J. Proteome Res., 2019



Fumagillin

Fumagillin suppresses HIV-1 infection of macrophages through the inhibition of Vpr activity

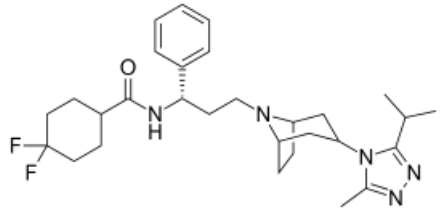
Nobumoto Watanabe, FEBS Lett., 2006



Trametinib

MEK1/2 selective allosteric inhibitor Trametinib reduces HIV-1 infectivity via the decrease in virion-incorporated ERK2 phosphorylation.

Takeo Dochi, Biochem Biophys Res Commun, 2018



Maraviroc

Maraviroc (MVC) is the only CCR5 antagonist currently approved by the FDA. MVC has been shown to be effective at inhibiting HIV-1 entry into cells and is well tolerated.

Shawna M Woollard, Drug Des Devel Ther., 2015

Thank you for your attention

This study is supported by the Russian Science Foundation grant № 19-75-10097

"Analysis of the interactions between HIV and human organism considering prescribed HIV/AIDS therapy"