SUPPORTING ONLINE MATERIAL FOR PREDICTION OF INTERACTIONS BETWEEN HIV-1 AND HUMAN PROTEINS BY INFORMATION INTEGRATION

OZNUR TASTAN[1], YANJUN QI[1], JAIME G. CARBONELL[1] AND JUDITH KLEIN-SEETHARAMAN[1, 2]

[1]School of Computer Science, Carnegie Mellon University 15213 and [2] Department of Structural Biology, School of Medicine, University of Pittsburgh, 15260, Pittsburgh, PA USA.

1. Methods Supplement

1.1. *Dataset*

The interactions between HIV-1 host proteins were retrieved from the NIAID database [1] before the November 17 2007 update for all HIV-1 proteins except env. Env was added in the database after the November update.

Table S1. Two exclusive groups of keywords were derived from the NIAID HIV-1, human interactions database [1]. Group 1 keywords represent the most likely direct physical interactions and Group 2 are those keywords referring to interactions that may be indirect.

Group 1 keywords	acetylated by, acetylates, binds, cleaved by, cleaves, degraded by, degrades, dephosphorylates, interacts with, methylated by, myristoylated by, phosphorylated by, phosphorylates, ubiquitinated by
Group 2 keywords	activated by, activates, antagonized by, antagonizes, associates with , cleavage induced by, causes accumulation of, co-localizes with, competes with, complexes with, cooperates with, decreases phosphorylation of, deglycosylates, depolymerizes, displaces, disrupts, downregulated by, downregulates, enhanced by, enhances, enhances phosphorylation of, enhances polymerization of, enhances release of, excludes, exported by, facilitated by, fractionates with, glycosylated by, imported by, inactivates, incorporates, induces, induces acetylation of, induces accumulation of, induces cleavage of, induces complex with, induces phosphorylation of, induces rearrangement of, induces release of, influenced by, inhibited by, inhibits, inhibits acetylation of, inhibits induction of, inhibits release of, inhibits release of, isomerized by, mediated by, modified by, modulated by, modulates, palmitoylated by, processed by, polarizes, promotes binding to, protects, recruited by, recruits, redistributes, regulated by, regulates, regulates import of, relocalized by, relocalizes, requires, sensitizes, sequesters, stabilizes, stimulated by, stimulates, synergizes with, transported by, upregulated by, upregulates

Table S2. The number of non-redundant interactions between HIV-1 proteins and human proteins according to the NIAID HIV-1, Human Interaction database [1]. The database architecture allows for reporting multiple interactions of the same protein pair, because multiple references may support an interaction. An interaction is classified as a Group 1 type of interaction if it is described by at least one of the Group 1 keywords and classified as Group 2, if otherwise. We did not use the gap p1protein in our analysis, since few annotations exist for this protein and only 2 Group 1 interactions and no Group 2 interactions are known.

HIV protein	**Number of HIV-1- Human Interactions**	
	Group 1 type	**Group 2 type**
Envelope gp41	37	118
Envelope gp120	195	336
Envelope gp160	54	121
Gag capsid	19	13
Gag matrix	39	37
Gag nucleocapsid	5	19
Gag p6	14	0
Gag pr55	15	32
Nef	71	119
Integrase	72	6
Protease	60	18
Reverse transcriptase	17	22
Rev	33	29
Tat	336	420
Vif	54	10
Vpr	35	134
Vpu	7	13
Total	1063	1454
Number of unique human proteins involved	721	914

1.2. *Random forest classifier*

The Random Forest (RF) is an ensemble of multiple independent decision trees. Each decision tree is constructed using a bootstrap sample of the training, and at each decision node in the tree, the best splitting feature is selected from randomly selected *m* features. To classify a new example, the feature vector is provided to all trees of the forest; each tree gives a decision, then the forest chooses the decision according to the majority vote of these trees. We used the Berkeley Random Forest package implementation of the RF classifier [2]. In our experiments, 200 trees were grown. To cope with the unbalanced class distribution of examples, the cost of mis-classifying a positive (‘interacting’) example is weighted more by a factor of *w* as compared to mis-classifying a negative (‘non-interacting’) example. The parameters *m* and *w* were tuned using only the training data in each cross-validation run; parameter values achieving the best mean average precision score (MAP) [3] were selected for training of the final model.

1.3. *Features*

We collected a total of 35 feature attributes, listed in Table S3. For all features, where appropriate, the missing values were substituted with the mean or median (for non-categorical and categorical attributes, respectively) of the non-missing values for that feature. This is a commonly used strategy for handling missing values in classification. The details of each feature and the biological sources of these features are discussed below. Some of our features are specific to the HIV-1, human protein pairs and contain information addressing the question whether the two proteins interact or not. Other features were related only to the human proteins in an HIV-1 relevant context. The latter features address the question whether a given human protein interacts with *any* HIV-1 protein. Table S3 also includes annotation regarding whether knowledge of the human-PPI network is used in the design of the feature.

Table S3. Feature set derived for prediction of interactions between HIV-1 and human proteins. We collected a total of 35 features. The first column indicates which of the features include information regarding the human protein interactome. The second column lists the name of the feature. The third column lists the number of features for each source. The fourth column, ‘Specific to’, describes whether the feature is specific to the HIV-1 protein pair, only to the human protein or only to the HIV-1 protein. The fifth column presents the percentage of pairs for which information is present (coverage). For gene expression features, the average coverage across the four gene expression data sets (Table S4) is given. In some of the features, coverage is 100 % as a result of the way the feature was encoded. For example, for the ELM-ligand feature, if the condition for the pair is not satisfied, the feature value for that pair takes a value of zero. In such cases, the percentage of non-zero elements is given. The features for which this applies are marked with * in the last column.

	Feature(s) Name	**Number of features**	**Specific to**	**Coverage**
	GO function similarity	1	Human, HIV-1 protein pair	65.4
	GO process similarity	1	Human, HIV-1 protein pair	63.3
	GO component similarity	1	Human, HIV-1 protein pair	66.7
Features that make use of human protein interactome	GO neighbor function similarity	1	Human, HIV-1 protein pair	42.6
	GO neighbor process similarity	1	Human, HIV-1 protein pair	45.3
	GO neighbor component similarity	1	Human, HIV-1 protein pair	45.3
	Post translational modification	1	Human, HIV-1 protein pair	40.0
	Degree	1	Human protein	45.3*
	Clustering coefficient	1	Human protein	20.0*
	Betweenness centrality	1	Human protein	31.6*
	Sequence similarity to human protein's neighbors	1	Human protein	45.3
	Pairwise sequence similarity	1	Human protein	100.0
	ELM, ligand feature	1	Human, HIV-1 protein pair	2.3*
	Gene expression features	4	Human protein	44.0
	Tissue distribution	1	Human protein	66.6
	HIV protein type features	17	HIV-1 protein	100.0

Provided below is detailed information regarding the collection and encoding of each feature.

1.3.1 ***GO similarity features:*** In the GO feature category, six features were derived. In each case**,** we applied the semantic similarity method G-SESAME [4] to measure the similarity between two proteins' annotations. In GO, each ontology is represented by a hierarchical directed acyclic graph (DAG), in which nodes are GO terms and edges are relationships between these terms. A child term may be an "instance" of its parents' term ("is-a" relationship) or a component ("part-of" relationship). To calculate the semantic similarity between two GO terms, G-SESAME compares the subgraphs of these GO terms (starting from the specific GO term ending in a root term). In doing so, G-SESAME not only considers the common ancestors the GO term pair have but also the location (closeness to the most specific term) and the relation type of the edges. Then, the GO semantic similarity between two proteins' annotation set is calculated by averaging the maximal similarity of each term to the other protein's annotation set. GO annotations were gathered from the European Bioinformatics Institute Gene Ontology Annotations database [5] and the ontologies were obtained from the GO consortium [6].

1.3.2 ***Graph properties of human proteins in the human interactome:*** This set of features was designed to encode the properties of the connections between human proteins in the human protein interactome. The human protein interactions can be described as an undirected graph $G = (V, E)$ with a set of vertices V(proteins) and a set of edges E(interactions). An edge e_{ij} connects vertex i with vertex j. We utilized three topological properties of a vertex: degree, clustering coefficient and betweenness centrality. The degree k_v of a vertex v is the number of its interaction partners. Clustering coefficient [7], on the other hand, measures the extent to which the protein's interaction partners are connected to each other and is defined as $C_v = 2\ n_v\ /\ k_v\ (k_v\ \text{-}1)$. In this equation, k_v is the number of interaction partners (degree) and n_v is the number of edges present between these neighbors. Clustering coefficient is defined for vertices with degree $k_v \geq 2$. Betweenness centrality is defined as in reference [8]; for a node it is calculated as the fraction of shortest paths between node pairs that pass through the node of interest. High betweenness centrality indicates that the protein has control over the information flow between other proteins in the network.

1.3.3 ***ELM-ligand feature:*** Short functional sequence motifs were downloaded from the Eukaryotic Linear Motif (ELM) database [9]. Based on the motif descriptions, we identified 109 ELMs, which mediate binding to a protein domain or a specific protein class. An example for a domain binding ELM is the sequence pattern PXXDY (ELM id: LIG_SH3_5), which is recognized by SH3 domains. Similarly, the XLX{0,1}[IVMP] sequence pattern (ELM id: LIG_CYCLIN_1) is found in the substrate recognition sites of cyclin interacting proteins. The feature is encoded as follows: if there exists an ELM motif in the HIV-1 protein's sequence and its ligand domain is present in the human binding partner, or the human binding partner belongs to the ligand class, the feature takes a non-zero value (see Finding conserved motifs, below), otherwise it is set to zero. Since there were no ELM binding motifs present in HIV-1 proteins and there were no ELM ligand domains present in the HIV-1 proteins, the mirror case was not applicable. The domain assignments for human proteins are obtained from the InterPro database [10].

The HIV-1 sequences mutate rapidly and some motifs are very short. To reduce the number of false positives that may result from these properties, we only consider a motif to match the HIV-1 sequences if the motif instance is conserved in multiple HIV-1 protein sequences and if the match of the ELM motif was classified as 'strong' or 'weak' according to their specificity (see 'Classification of motifs', below). More specifically, let *i* be the HIV-1 protein *i,* the set of HIV-1 sequences Q_i and the set of ELM motif's ligand class, L_m, the feature value f_{elm} for pair *i* and *j* will be,

$$\mathrm{f}_{\mathrm{elm}}(i,j)=\begin{cases}1 & \text{if } \exists\, m \text{ such that } m \text{ is conserved in } \mathrm{Q}_i \\ & \text{and } j\in L_m \text{ and } m \text{ is strong} \\ 0<q<1 & \text{if } \exists\, m \text{ such that } m \text{ is conserved in } \mathrm{Q}_i \\ & \text{and } j\in L_m \text{ and } m \text{ is weak} \\ 0 & \text{otherwise}\end{cases} \qquad (1)$$

Classification of motifs: The classification of motifs as either strong or weak is based on motif specificity. If the probability that a random sequence would match the motif pattern is low, the motif is highly specific ('strong'), otherwise it is not specific. An example of a non-specific motif would be the sequence pattern 'AA' because Alanine is a frequently observed amino acid in protein sequences and the motif is short, thus the chances that a random sequence would match this motif is high. Motif specificity is calculated based on the assumption that the distribution of amino acids in each position of a random sequence is independent and identically distributed as described by Nevill-Manning et al. [11]. The observed distribution of each amino acid in the redundant sequence database of UniRef50 [12] was used to estimate the probability of an amino acid occurring in a position. Motifs with specificity ≤0.001 are considered as strong; whereas others are considered weak. 0.001 is a stringent threshold. The parameter *q* for weak motifs (Eq.1) is tuned on the training data of each cross-validation step. *q* was assigned a value of 1 if the motif is strong, and 0 in cases where no motif was found, and weak motifs take values in between. The parameter was varied and the best value was chosen as the one that achieved the best chi-square (see section 'Feature selection').

Finding conserved motifs: HIV-1 sequences mutate rapidly; hence, the sequences that code for the same proteins show high variability in multiple regions. In order to minimize false positive motif matches, we consider a motif a match in the HIV-1 protein sequence only if it is conserved. The conserved ELMs are identified as follows: First, the motif pattern is identified in all HIV-1 protein sequences. Then a block of motif instances, whose positions are close in the multiple sequence alignment, is searched for. To consider an ELM as conserved, we require at least λ fraction of all the protein sequences to have a motif instance within five residues of the other sequences' motif instances. The five residue window is used to accommodate possible shift errors in the alignment. The degree of conservation required, λ, is tuned on the training data of each cross-validation step. The multiple sequence alignments of HIV-1 proteins were retrieved from the Los-Alamos HIV sequence database

[13]. In multiple alignments, only HIV-1 sequences that belong to group M clade were used [14]. These are translated protein sequences, thus have alignments of nine translated genes. In cases where an HIV-1 gene encodes multiple HIV-1 protein products, the alignments were split such that they will cover only the relevant protein coding region.

1.3.4 ***Differential gene expression features:*** Using three gene expression datasets retrieved from the Gene Expression Omnibus [15], four features were built. The sources and details of these datasets are listed in Table S4. The features encode the log fold change in expression levels. Let $<E^{+}>$ and $<E^{-}>$ denote the average expression levels of genes in HIV-1 infected and uninfected samples, respectively, in their original scales, the log fold change is calculated as $\log(<\mathrm{E}^{+}>/<\mathrm{E}^{-}>)$. In the case of time series experiments, samples at each time point are compared individually. The fold change of the time point, where the maximum absolute fold change occurs, is then used as the feature value.

Table S4. Gene expression data sets used as features. Gene expression data sets are derived from the Gene Expression Omnibus (GEO) of NCBI [15] available at http://www.ncbi.nlm.nih.gov/geo/. The dataset ids are given in the first column, which can be queried in GEO. The second column describes the dataset. In the third column, samples are compared. The fourth column list the number of features derived from the respective dataset. The reference reporting the gene expression study is given in the last column.

GDS ID	Description	Compared Samples	Number of features derived	Reference
GDS2649	CD4+ and CD8+ T cells from HIV patients at different clinical stages and rates of disease progression. 40 samples: 20 for CD4+ and 20 CD8+; four different disease states in each: Chronic-progressive(5), early infection(5), uninfected(5) and non-progressive(5).	4 different disease states reduced into two as was done in the original paper. Two comparisons done for CD4+ and CD8+ separately, the number of samples are given in the parenthesis: 1. [Uninfected(5) + Non-progressive(5)] CD4+ samples vs. [Chronic-progressive(5) + Early infection(5)] CD4+ samples 2. [Uninfected (5) + Non-progressive (5)] CD8+ vs. [Chronic- progressive(5) + Early infection(5)]CD8+	2	[16]
GDS1726	Analysis of brain frontal cortex of HIV-seropositive patients with HIV encephalitis (HIVE). 2 disease state. HIV-seropositive patients with HIV encephalitis. 28 samples: 16 disease and 12 control samples.	One comparison: HIV encepthalitis (16) *vs.* uninfected control samples (12).	1	[17]
GDS171	Gene expression profile of proliferating normal peripheral blood mononuclear cells (PBMC) infected with HIV type 1 RF. 3 infected and 3 uninfected for each time samples for each time points: t=0,12,24,48,72 hours	5 comparisons conducted: At each time point infected (3) *vs.* uninfected (3) samples are compared.	1	[18]

1.3.5 ***Tissue feature:*** If a protein is not expressed in those tissues that HIV-1 can infect, the likelihood of an interaction between an HIV-1 protein and this protein is very small. To use this information for prediction, a tissue similarity feature is encoded as follows. If a human gene's annotated tissues are enriched with HIV susceptible tissues, the feature value is high and low if it is not. We varied the feature encoding (binary, ratio of common tissues, and what to encode if the human gene is expressed ubiquitously) and selected the best encoding of this feature based on chi-square ranking of the feature on the training data set. The tissues expressing the human proteins were collected from the Human

Protein Reference Database (HPRD) [19] and the Human Proteinpedia (HUPA) website [20]. 13,920 human genes are annotated with at least one tissue according to HPRD and HUPA. The tissues susceptible to HIV-1 infection were obtained from Levy *et al.*[21].

1.3.6 ***Sequence similarity features:*** For each pair, two sequence similarity features are utilized, pairwise sequence similarity and similarity to human proteins' human protein interaction partners. The motivation behind the sequence similarity feature is the fact that homo-oligomers including homo-dimers are frequent in protein structures; similarly two protein structures with similar sequences, possibly similar structures might interact as well. Pairwise sequence similarity is computed for each pair using the BLASTP package [22]. The best hit subsequence's -log(e-value) is used as the similarity measure, where e-value is the expected value of the alignment score to be found merely by chance [22]. In calculating similarity between the HIV-1 protein and the human protein's interaction partners, the pairwise similarities are first calculated then the maximum similarity score is chosen among these.

1.3.7 ***Posttranslational modification similarity to neighbor:*** Some protein interactions require the protein(s) to be in a certain posttranslationally modified state. For such cases, the HIV-1 protein will require to mimic the posttranslational modification (PTM) of the human protein's interaction partner. For example, the N-terminal myristoylation domain of Nef is highly conserved and plays an important role in interacting with calmodulin; one of calmodulin binding partners is the human protein NAP-22/CAP23, which is also myristoylated [23]. The feature is encoded as binary such that for a pair, the feature takes a value of 1 if at least one of the human protein's human protein binders shares at least one common PTM with the HIV-1 protein. To this end, we collected the HIV-1 proteins' PTMs from the literature. Experimentally verified posttranslational modifications of human proteins were obtained from three different databases, HPRD [19], HUPA [20] and database of Post Translational Modifications (dbPTM) [24].

1.3.8 ***Human protein interactome:*** In the features where human inter-PPI information is involved, the human proteome is based on the interactions reported in the Human Protein Reference Database (HPRD) [19, 25] Release 7, which includes 37.083 interactions involving 9460 human proteins.

1.4. *Feature selection*

In features where multiple encodings are possible, such as the degree of conservation, λ, we employed chi-square feature selection [26] to identify the best means of encoding. Using only the training data in each cross-validation step, the parameter combinations were chosen such that the best chi-square score is achieved.

2. Results Supplement

2.1. *Predictions*

Predicted protein pairs with scores above a random forest (RF) score of zero can be accessed in form of an excel file at the Supplementary Information link http://www.cs.cmu.edu/~oznur/hiv/hivPPI.html (click on the predictions link). The excel file includes protein pairs ranked according to their RF score. The higher the score, the more likely is the interaction. The columns in the table contain the following information:

Column A ("HIV-1 protein"): the HIV-1 protein involved in the predicted interaction.
Column B ("Human partner Entrez gene id"): NCBI Entrez Gene ID of the human protein involved in the predicted interaction.
Column C ("Human partner gene symbol"): Gene symbol of the human gene.
Column D ("Human partner official name"): Gene name of the human gene according to NCBI Entrez Gene database.
Column E (" Random forest score"): The random forest score assigned to the predicted pair by the model.
Column F ("Is the interaction a Group 1 interaction?"): If the interaction is reported in NIAID database and is reported by at least one of the Group1 keywords (see Supplementary Table S1), the keywords reporting the interaction are listed. If the interaction is not classified by a Group1 keyword, the cell will contain "No".
Column G (**"Is the interaction a Group 2 interaction?"):** If the interaction is reported in NIAID and is one of the Group2 interactions (not reported by any of the Group1 keywords (see Supplementary Table S1). The keywords reporting the interaction is listed.
Column H ("Prediction Type"): If the predicted pair is one of the Group1 interactions, the cell will list Group1, if it is a Group2 interaction, they will list Group2 and if it is not Group1 or Group2, it is listed Novel.
Column I ("Is the human gene detected in Virion?"): If any of the human gene's gene products is reported to be hijacked into the HIV-1 virion in the literature the cell lists "Yes", otherwise "No".
Column J ("Is the human gene reported in Brass *et al.* siRNA screen?"): If the human gene is reported in the siRNA study, the cell lists "Yes", otherwise "No".
Column K ("Does the human partner interact with a siRNA reported gene?"): If the human gene is an interactor of any of the siRNA genes, the cell lists those genes, if not it is listed as "No". The human protein interactome is based on the Human Protein Interaction Database [19].

2.2. *GO enrichment*

Table S5. Enriched GO molecular function in the unique set of human proteins that involve novel and Group 2 predicted interactions**.** The enrichment is calculated for the unique human proteins of these predicted interactions, and the *p*-value is adjusted via Bonferroni correction for multiple hypothesis testing.

GO term ID	**GO term name**	***p*-value**
GO:0030528	transcription regulator activity	3.52e-38
GO:0005515	protein binding	2.97e-29
GO:0005488	binding	9.33e-20
GO:0008134	transcription factor binding	7.75e-18
GO:0003677	DNA binding	1.83e-13
GO:0004879	ligand-dependent nuclear receptor activity	1.85e-12
GO:0005057	receptor signaling protein activity	1.02e-08
GO:0016740	transferase activity	5.64e-08
GO:0032393	MHC class I receptor activity	5.83e-08
GO:0003676	nucleic acid binding	1.24e-07
GO:0016563	transcription activator activity	6.98e-07
GO:0060089	molecular transducer activity	2.60e-06
GO:0051427	hormone receptor binding	5.92e-05
GO:0016772	transferase activity, transferring phosphorus-containing groups	8.19e-05
GO:0008565	protein transporter activity	1.53e-04
GO:0047485	protein N-terminus binding	1.79e-04
GO:0004697	protein kinase C activity	6.05e-04
GO:0003682	chromatin binding	7.38e-04
GO:0000166	nucleotide binding	7.53e-04

Table S6. Enriched GO molecular processes in the unique set of human proteins that involve novel and Group 2 predicted interactions. The enrichment is calculated for the unique human proteins of these predicted interactions, and the *p*-value is adjusted via Bonferroni correction for multiple hypothesis testing.

GO term ID	GO term name	*p*-value
GO:0002376	immune system process	2.03e-54
GO:0051704	multi-organism process	2.11e-45
GO:0050896	response to stimulus	8.12e-44
GO:0043170	macromolecule metabolic process	3.83e-24
GO:0008150	biological_process	5.02e-24
GO:0019882	antigen processing and presentation	1.87e-19
GO:0032502	developmental process	6.36e-19
GO:0048518	positive regulation of biological process	1.73e-17
GO:0044419	interspecies interaction between organisms	2.47e-16
GO:0010467	gene expression	4.30e-16
GO:0065007	biological regulation	3.47e-14
GO:0016265	death	2.46e-11
GO:0048519	negative regulation of biological process	1.44e-10
GO:0043283	biopolymer metabolic process	1.01e-09
GO:0006139	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	5.69e-09
GO:0010604	positive regulation of macromolecule metabolic process	6.15e-09
GO:0007049	cell cycle	2.67e-08
GO:0007242	intracellular signaling cascade	5.81e-08
GO:0006366	transcription from RNA polymerase II promoter	2.18e-07
GO:0007167	enzyme linked receptor protein signaling pathway	9.32e-07
GO:0006793	phosphorus metabolic process	1.69e-06
GO:0044238	primary metabolic process	1.80e-05
GO:0007154	cell communication	1.86e-04
GO:0008283	cell proliferation	3.05e-04
GO:0030522	intracellular receptor-mediated signaling pathway	6.29e-04
GO:0009058	biosynthetic process	1.76e-03
GO:0048002	antigen processing and presentation of peptide antigen	1.96e-03
GO:0043284	biopolymer biosynthetic process	2.16e-03
GO:0050793	regulation of developmental process	3.15e-03
GO:0016310	phosphorylation	4.59e-03
GO:0016043	cellular component organization and biogenesis	4.65e-03

Table S7. Enriched GO cellular components in the unique set of human proteins that involve novel and Group 2 predicted interactions. The enrichment is calculated for the unique human proteins of these predicted interactions, and the *p*-value is adjusted via Bonferroni correction for multiple hypothesis testing.

GO term ID	**GO term name**	***p*-value**
GO:0032991	macromolecular complex	2.07e-45
GO:0031974	membrane-enclosed lumen	2.76e-35
GO:0005886	plasma membrane	1.36e-24
GO:0042611	MHC protein complex	1.26e-18
GO:0005829	cytosol	1.30e-17
GO:0043226	organelle	1.19e-13
GO:0044459	plasma membrane part	1.68e-13
GO:0005634	nucleus	2.11e-10
GO:0043234	protein complex	6.48e-10
GO:0005575	cellular_component	1.47e-08
GO:0043227	membrane-bounded organelle	6.19e-08
GO:0005623	cell	6.69e-07
GO:0044424	intracellular part	5.66e-05
GO:0005694	chromosome	7.88e-05

2.3 *Feature importance*

The precision vs. recall curve of the RF model in which protein type features were excluded during training is compared to the model trained with the full feature set in Figure S1.

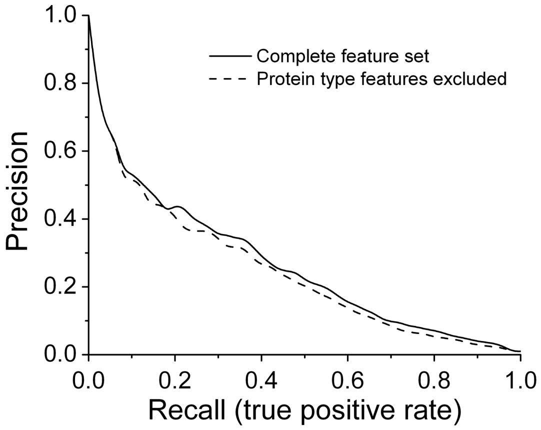

Figure S1. The precision vs. recall curves of the models trained with and without protein type features.

9

2.4 AUC scores in false positive range

Table S8. AUC scores computed in false positive range.

	AUC0.1	AUC0.05	AUC0.01	AUC0.001
Avg	0.6092	0.4958	0.2374	0.0527
Std	0.0183	0.0218	0.0235	0.0125

References:

1. *NIAID HIV-1, Human Protein Interaction Database*. [http://www.ncbi.nlm.nih.gov/RefSeq/HIVInteractions/].
2. L. Breiman, *Random Forests*, in *Machine Learning*. 2001. p. 5-32.
3. R.R. Baeza-Yates and B.d.A. Neto, *Modern information retrieval*. 1999, New York Harlow, England: ACM Press; Addison-Wesley.
4. J.Z. Wang *et al.*, *Bioinformatics* **23**, 1274-81 (2007).
5. C. Brooksbank, G. Cameron and J. Thornton, *Nucleic Acids Res* **33**, D46-53 (2005).
6. M. Ashburner *et al.*, *Nat Genet* **25**, 25-9 (2000).
7. D.J. Watts and S.H. Strogatz, *Nature* **393**, 440-2 (1998).
8. L.C. Freeman, *Sociometry* **40**, 35-41 (1977).
9. P. Puntervoll *et al.*, *Nucleic Acids Res* **31**, 3625-30 (2003).
10. R. Apweiler *et al.*, *Nucleic Acids Res* **29**, 37-40 (2001).
11. C.G. Nevill-Manning, T.D. Wu and D.L. Brutlag, *Proc Natl Acad Sci U S A* **95**, 5865-71 (1998).
12. C.H. Wu *et al.*, *Nucleic Acids Res* **34**, D187-91 (2006).
13. *Los Alomos Database*. [http://www.hiv.lanl.gov/].
14. Los Alamos National Security. *HIV Sequence Compendium.* (2005).
15. T. Barrett *et al.*, *Nucleic Acids Res* **35**, D760-5 (2007).
16. M.D. Hyrcza *et al.*, *J Virol* **81**, 3477-86 (2007).
17. E. Masliah *et al.*, *J Neuroimmunol* **157**, 163-75 (2004).
18. M.T. Vahey *et al.*, *AIDS Res Hum Retroviruses* **18**, 179-92 (2002).
19. S. Peri *et al.*, *Genome Res* **13**, 2363-71 (2003).
20. S. Mathivanan *et al.*, *Nat Biotechnol* **26**, 164-7 (2008).
21. J.A. Levy, *HIV And the Pathogenesis of AIDS*. Third ed. 2007, Washington, DC: ASM Press. 82-91.
22. S.F. Altschul *et al.*, *J Mol Biol* **215**, 403-10 (1990).
23. M. Matsubara *et al.*, *Protein Sci* **14**, 494-503 (2005).
24. T.Y. Lee *et al.*, *Nucleic Acids Res* **34**, D622-7 (2006).
25. G.R. Mishra *et al.*, *Nucleic Acids Res* **34**, D411-4 (2006).
26. Y. Yang and J.O. Pedersen. *14th International Conference on Machine Learning*. 1997.