Patterns of amino acid conservation in human and animal immunodeficiency viruses

Olga S. Voitenko1,2*, Andi Dhroso3, Anna Feldmann1,2, Dmitry Korkin3 and Olga V. Kalinina1,*

1Department for Computational Biology and Applied Algorithmics, Max Planck Institute for Informatics, Campus E1 4, Saarbrücken 66123, Germany, 2Graduate School for Computer Science, Saarland University, Campus E1 3, Saarbrücken 66123, Germany, and 3Department of Computer Science and Bioinformatics and Computational Biology Program, Worcester Polytechnic Institute, Worcester, MA 01609, USA

*To whom correspondence should be addressed.

Abstract

Motivation: Due to their high genomic variability, RNA viruses and retroviruses present a unique opportunity for detailed study of molecular evolution. Lentiviruses, with HIV being a notable example, are one of the best studied viral groups: hundreds of thousands of sequences are available together with experimentally resolved three-dimensional structures for most viral proteins. In this work, we use these data to study specific patterns of evolution of the viral proteins, and their relationship to protein interactions and immunogenicity.

Results: We propose a method for identification of two types of surface residues clusters with abnormal conservation: extremely conserved and extremely variable clusters. We identify them on the surface of proteins from HIV and other animal immunodeficiency viruses. Both types of clusters are overrepresented on the interaction interfaces of viral proteins with other proteins, nucleic acids or low molecular-weight ligands, both in the viral particle and between the virus and its host. In the immunodeficiency viruses, the interaction interfaces are not more conserved than the corresponding proteins on an average, and we show that extremely conserved clusters coincide with protein–protein interaction hotspots, predicted as the residues with the largest energetic contribution to the interaction. Extremely variable clusters have been identified here for the first time. In the HIV-1 envelope protein gp120, they overlap with known antigenic sites. These antigenic sites also contain many residues from extremely conserved clusters, hence representing a unique interacting interface enriched both in extremely conserved and in extremely variable clusters of residues. This observation may have important implication for antiretroviral vaccine development.

Availability and Implementation: A Python package is available at https://bioinf.mpi-inf.mpg.de/publications/viral-ppi-pred/

Contact: voitenko@mpi-inf.mpg.de or kalinina@mpi-inf.mpg.de

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

Viral proteins participate in a multitude of interactions, including those with other components of the viral particle and with molecules in the host cell. High-throughput analysis of protein–protein interactions between various viruses and their hosts revealed many unexpected links between viral proteins and the host metabolic machinery (Jäger et al., 2011; Pichlmair et al., 2012). Viruses with RNA-encoded genome have a characteristic high rate of mutation, due to the lack of the proof-reading mechanisms in their RNA-dependent RNA polymerases and reverse transcriptases (Cuevas et al., 2015). The high rate of mutations enables viruses to evade the host immune response (Drake et al., 1998). This creates two seemingly contradicting evolutionary forces: a negative selection pressure to maintain the virus–host interactions and a positive selection pressure to keep viral proteins as variable as possible in order to avoid recognition by the immune system. In this work, we sought to explore how these tendencies shape the patterns of amino acid conservation of the surface of viral proteins.

Conserved residues are known to play an important role in maintaining the structure of various proteins, see e.g. Shimokawa et al.
It has been previously shown that the buried core of the protein is more highly conserved than its surface (Rodionov and Blundell, 1998). Protein–protein interaction interfaces (Guharoy and Chakrabarti, 2005; Valdar and Thornton, 2001) and active sites (Bartlett et al., 2002; Guharoy and Chakrabarti, 2010) have also been reported to be more conserved than the protein surface on average. These analyses have not been specifically focused on viral proteins, although there is emerging evidence that intra-viral protein interactions also tend to contain many well-conserved residues (Warren et al., 2013). Similar to other pathogens, viruses participate in the arms race with the host immune system, which may result in positive selection pressure on the virus–host interaction interfaces (Sirone et al., 2013; Trivedi and Wang, 2014; Woolhouse et al., 2002). At the same time, unlike proteins of the immune system, viral receptor proteins do not vary much among the potential hosts, so the interaction sites with them need to be conserved, whereas the antigenic sites can be expected to evolve faster than the viral protein on average.

In this work, we introduce a concept of extremely conserved and extremely variable clusters as compact patches of amino acid residues on a protein surface that collectively display either unusually high (for extremely conserved) or unusually low (for extremely variable) conservation in a family of sequences. We present a new method that allows to simultaneously identify extremely conserved and extremely variable clusters in any protein, for which an experimentally resolved or a modeled three-dimensional structure is available. We apply it here to studying conservation pattern of viral proteins, since they present a unique case of a fast and flexibly evolving system, engaged in complex interactions with each other and with the host cell.

Because of their medical importance, we focus on human and animal immunodeficiency viruses, from the genus Lentivirus of the Retroviridae family. The human immunodeficiency virus (HIV) causes AIDS and is one of the major health hazards, being responsible for 1.2 million deaths worldwide in 2014 (http://www.who.int/). HIV-1 is the globally spread type of HIV, which causes the majority of infections, while HIV-2 is less prevalent. HIV has been a subject of intensive research for more than three decades, and in this work, we benefit from the data on protein sequences, structures and interactions collected for this virus. Other viruses from the same genus, such as the simian immunodeficiency virus (SIV) and feline immunodeficiency virus (FIV), infect other mammals and usually do not cause any symptoms (VandeWoude and Apestli, 2006). The protein sequences from different immunodeficiency virus species show significant divergence (55 ± 13% sequence identity between HIV-1 and HIV-2, 53 ± 12% between HIV-1 and SIV, 38 ± 13% between HIV-1 and FIV), but their three-dimensional structures show remarkable similarity. Compared with HIV-1, little is known about interactions of the proteins from HIV-2, SIV and FIV.

We observe a strong correlation between extremely conserved clusters and protein interaction interfaces with other proteins, nucleic acids and low molecular-weight ligands. Particularly, these clusters overlap well with hotspots of protein–protein interactions, i.e. groups of residues that contribute most to the energy of a protein–protein interaction. In this work, we introduce and explore in detail the concept of extremely variable clusters that comprise positions of very low conservation, which may be an indication of positive selection. We show that in the HIV-1 envelope protein, extremely variable clusters are located in antigenic sites, i.e. sites targeted by neutralizing antibodies. Interestingly, the antigenic sites also contain many extremely conserved clusters. We also observe significant differences in the distribution of conserved and variable clusters among human and animal immunodeficiency viruses, but it is unclear whether these differences are caused by true biological differences or by an uneven sequence sampling.

2 Methods

The study comprises three principal steps: (1) data acquisition and processing, which includes retrieval, filtering and alignment of protein sequences and retrieval of corresponding protein structures; (2) calculation of conservation scores for each alignment position; and (3) identification of extremely conserved and extremely variable tight clusters.

2.1 Data acquisition and processing

All sequences of HIV-1 proteins were downloaded from the Los Alamos HIV sequence database (http://hiv.lanl.gov). All fragments shorter than 30% of the length of the corresponding protein in the HIV-1 HXB2 reference strain were removed. We further refer to this set as LANL dataset. Sequences of HIV-2 and SIV proteins were also downloaded from the Los Alamos HIV sequence database (HIV-2 and SIV datasets). For FIV proteins, sequences were downloaded from the Uniprot database (Consortium, 2015) using BLAST search with the corresponding HIV-1 protein among sequences from Feline immunodeficiency virus (taxon identifier: 11674). Only first 1000 hits were retained (FIV dataset). The sequences were then filtered to be not more than 90% identical using CD-HIT (Fu et al., 2013) and aligned using HIV-align (Eddy, 1998). A three-dimensional structure from the Protein Data Bank (Berneman et al., 2000) was chosen for every HIV-1 protein based on the sequence coverage and resolution. For all viruses, structures of the corresponding HIV-1 proteins were used. Additionally, we considered HIV-1 sequences from the EuResist database (Rosen-Zvi et al., 2008) (EuResist dataset) and from two transmission chains (Leitner et al., 1996; Lemey et al., 2005), employing the same filtering protocol (Table 1). The latter two datasets were aligned with MAFFT (Katoh and Standley, 2013).

2.2 Conservation of individual positions

For each alignment position $i$, the conservation score was calculated using the measure suggested in Valdar (2002)

$$c_i = (1 - t(i))^a(1 - r(i))^b(1 - g(i))^c$$

where $t(i)$ represents the diversity of amino acid residues at position $i$ computed as the position Shannon entropy; $r(i)$ represents stereochemical diversity computed as the average distance of each amino acid residue, expressed as values of the BLOSUM62 substitution matrix, from all other residues of the column, normalized by the average pairwise distance among all residues in the column (for details, see Thompson et al., 1997); $g(i)$ is the fraction of gaps at position $i$. The parameters $a$, $b$ and $c$ are set to 1, 1/2 and 2, respectively (as suggested in Valdar, 2002).

2.3 Clustering

We set out to identify two sets of clusters of amino acid residues on the surface of the proteins—those of extremely conserved residues and those of extremely variable residues—using the same algorithm. The core of the method is the tight clustering procedure (Tseng and Wong, 2005), which is based on the idea that clusters that recurrently appear in subsets of the data constitute distinct tight formations in the whole dataset. The procedure thus involves multiple iterations ($B$ times) of clustering a randomly drawn fraction of the data of size $S$, where clusters that overlap with Jaccard coefficient
For each protein, sequence counts correspond to the initial dataset size before filtering.

The two target sets of clusters were identified by using two distance metrics that take into account both physical distance between residues and their conservation. For computations, only the surface residues were considered, defined as having relative solvent accessible area greater than 16% (Rost and Sander, 1994) in the corresponding protein structures, computed with NACCESS (Hubbard and Thornton, 1993). Two surface residues with their closest atoms less than 5 Å apart were considered contacting, and the surface distance between them is defined as the Euclidean distance between these two atoms. To compute the surface distance between non-contacting residues, first the network of all contacting residues is constructed with RNALyzer (Doncheva et al., 2011), and then the shortest path was found between each pair of non-contacting residues.

We modified the surface distance $d_{ij}$ between every pair of surface residues $(i, j)$ to take into account their conservation ($c_i, c_j$), calculated as described above:

$$d_{ij}^{\text{com}} = d_{ij} \left( \frac{2}{c_i + c_j} - \frac{1}{2} \right),$$

$$d_{ij}^{\text{var}} = d_{ij} \left( \frac{2}{2 - c_i - c_j} - \frac{1}{2} \right).$$

This distance measure contains an extra term that, in case of $d_{ij}^{\text{com}}$, redefines the distance between $i$ and $j$ to be smaller than their physical surface distance, if $i$ and $j$ are well conserved, and greater than their physical surface distance, if they are poorly conserved. The opposite holds for $d_{ij}^{\text{var}}$. This modification of the distance measure for clustering leads to the identification of tightly clustered patches of conserved residues for $d_{ij}^{\text{com}}$, and tightly clustered patches of variable residues for $d_{ij}^{\text{var}}$. Similarly, the procedure is stable with respect to the weight of the conservation-dependent term in the formulation of the modified distances.

### 3 Results

#### 3.1 Overlap with observed protein interfaces in the LANL dataset

To test the potential biological function of the identified conservation clusters, we have compared them with the interaction interfaces of HIV-1 proteins and other viral and human proteins, as well as nucleic acids and low molecular-weight ligands (Figure 1 and Supplementary Material). We define interface residues as those located less than 5 Å away from any of these molecules in any experimentally resolved three-dimensional structure of the corresponding viral protein. It must be noted that the average conservation of these interfaces does not significantly differ from the average conservation of the surfaces of the viral proteins (25 ± 20% versus 24 ± 19%), so we cannot expect extremely conserved clusters to cover them fully. Hence the positive predictive value (PPV, equals the fraction of the correctly identified interface residues in the clusters) and specificity (the fraction of non-interface residues correctly identified as being outside of tight clusters) are relevant measures for assessing the tendency of the clusters to lie on protein interaction interfaces (Table 2). The average PPV of the prediction exceeds 70%, and the average specificity is over 85%. For all HIV-1 proteins both these measures are greater than 50%, meaning that the predicted clusters are in a very good agreement with the observed interaction interfaces.

It is important to note that it is not guaranteed that all the protein interaction sites are observed in experimentally resolved three-dimensional structures. Experimental data suggests that the HIV-1 proteins participate in many more interactions, particularly with host proteins (Jäger et al., 2011), for which experimentally resolved structures of the corresponding complexes are not available. This for example is the case for the HIV-1 auxiliary proteins, of which Nef and Tat have been analyzed in our study. Nef presents the lowest PPV and specificity values among all proteins. It plays an important role in viral replication in vivo and enhances the infectivity of the viral progeny by mechanisms that are not fully understood (Basmaciogullari and Pizzato,
2014). On the surface of Nef, we observe large extremely conserved and extremely variable clusters that may constitute novel binding sites with unknown partners implicated in these processes (Fig. 1e).

Protein–protein interaction interface residues constitute 82% of all interaction interface residues considered in this study. However, not all the residues of protein–protein interaction interfaces are equally important for protein binding (Keskin et al., 2005; Sharma et al., 2005).

Table 2. Comparison of identified extremely conserved and extremely variable clusters to experimentally observed protein interaction interfaces with other viral and host proteins, nucleic acids and low molecular-weight ligands, as well as protein–protein interaction hotspots ("n/a": no prediction from Robetta (Kim et al., 2004))

<table>
<thead>
<tr>
<th>Protein</th>
<th>PDB identifier</th>
<th>Relative to observed interaction interfaces</th>
<th>Relative to predicted protein-protein interaction hotspots (Robetta)</th>
<th>Percentage of clustered residues among all surface residues (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>gp120</td>
<td>3J5M</td>
<td>59.26</td>
<td>10.81</td>
<td>90.35</td>
</tr>
<tr>
<td>Integrase</td>
<td>4OVL</td>
<td>72.73</td>
<td>15.69</td>
<td>93.75</td>
</tr>
<tr>
<td>Protease</td>
<td>3PWR</td>
<td>66.67</td>
<td>19.51</td>
<td>87.10</td>
</tr>
<tr>
<td>Reverse transcriptase</td>
<td>4PUO</td>
<td>60.53</td>
<td>15.65</td>
<td>92.96</td>
</tr>
<tr>
<td>p17 (matrix)</td>
<td>1HIW</td>
<td>63.64</td>
<td>16.28</td>
<td>89.47</td>
</tr>
<tr>
<td>p7 (nucleocapsid)</td>
<td>1F6U</td>
<td>100.00</td>
<td>14.89</td>
<td>100.00</td>
</tr>
<tr>
<td>p24 (capsid)</td>
<td>3MGE</td>
<td>64.71</td>
<td>12.50</td>
<td>89.09</td>
</tr>
<tr>
<td>Nef</td>
<td>4ORZ</td>
<td>53.36</td>
<td>19.23</td>
<td>72.41</td>
</tr>
<tr>
<td>Tat</td>
<td>3MI9</td>
<td>94.74</td>
<td>41.86</td>
<td>50.00</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>70.87</td>
<td>18.49</td>
<td>85.01</td>
</tr>
</tbody>
</table>

\(^a\)PPV = true positives/(true positives + false positives).

\(^b\)Sensitivity = true positives/(true positives + false negatives).

\(^c\)Specificity = true negatives/(true negatives + false positives).
et al., 2002). Hence, a notion of protein interaction hotspot has been previously introduced (Morrow and Zhang, 2012) to describe the residues, whose mutation greatly contributes to the decrease of binding energy between the two proteins. In the absence of the possibility to identify the protein interaction hotspots for HIV-1 protein experimentally, we have predicted them utilizing Robetta (Kim et al., 2004), an established tool in this field. Our clustering procedure identifies only a fraction of the hotspots as components of extremely conserved clusters; however, there is a clear enrichment of interaction hotspots in these clusters (Table 2). The specificity of hotspot prediction is very high for all HIV-1 proteins, meaning that the majority of the identified clusters consist of residues, whose mutation is likely to cause a large energetic impact on the protein–protein interaction.

The rest 18% of protein interaction interfaces comprise interaction with RNA and low molecular weight ligands, predominantly antiretroviral active compounds. We have separately analyzed these interaction interfaces (Supplementary Table S1). For reverse transcriptase, that has extensive RNA contact and many known inhibitors, sensitivity of the prediction increases by 5.65%. For capsid, it also increases by 4.20%, indicating that assembly inhibitors tend to bind in extremely conserved sites of protein–protein interaction interfaces.

3.2 Conserved and variable clusters on the virus–antibody interaction interfaces
While most interaction interfaces of HIV-1 are enriched in clusters of extremely conserved residues (Fig. 1), antigenic sites on the viral envelope protein gp120 present a special case. This protein is directly exposed to the pressure from the natural response of the host immune system, so binding sites for neutralizing antibodies, that have been identified so far, tend to evolve quickly (Walker et al., 2011), rendering most of the antibodies ineffective for potential therapy. Thus, we specifically investigated the distribution of extremely conserved and extremely variable clusters in these interfaces. We define four classes of antigenic sites for antibodies binding in different regions of gp120, encompassing 79 residues in total: associated with the CD4-binding site (Zhou et al., 2010; Wu et al., 2011); proximal to V1/V2 variable loops (McLellan et al., 2011); proximal to V3 loop (Walker et al., 2011); and other predicted epitope residues (West et al., 2013). Of the total 27 amino acid positions falling into either extremely conserved or extremely variable clusters, 15 are located in one of these antigenic sites, which makes 18.99% of their total size. This presents an enrichment, compared with 10.31% of all residues of gp120 falling into these clusters.

The largest fraction of neutralizing antibodies bind in or in close proximity to the binding sites of the receptor protein CD4, forming a large antigenic site of 47 amino acid residues. These include four residues from extremely conserved and five residues from extremely variable clusters. V1/V2-proximal and V3-proximal antigenic sites consist of 5 and 10 residues and include one and four residues from extremely conserved clusters, respectively. The fourth site, consisting of 17 residues, includes two residues from extremely variable clusters, and no extremely conserved residues (Fig. 2). Thus, the antigenic epitopes tend to overlap both with extremely conserved and with extremely variable clusters. Due to their characteristic, conserved clusters on the surface of gp120 might rather contain antigenic sites for neutralizing antibodies that bind a broad spectrum of viral strains (the so-called broadly neutralizing antibodies), while sites in the variable clusters are more likely to be targeted by narrow neutralizing antibodies.

3.3 Other sequence sets and immunodeficiency viruses
We additionally studied the distribution of extremely conserved and extremely variable clusters in other datasets. The overlap of the identified conserved and variable clusters in the LANL and the EuResist datasets is large: the corresponding Jaccard coefficient (the ratio between the size of the overlap and the size of the union of two sets (Tan et al., 2005)) for all identified clusters in all proteins is 0.25. The overlap of the identified clusters in the EuResist dataset with either protein interaction interfaces or protein–protein interaction hotspots is not significantly different. Interestingly, the variable clusters have a higher overlap than the conserved clusters with Jaccard coefficients 0.28 and 0.20, respectively. Proteins, for which more sequences are available in EuResist, such as protease, demonstrate yet higher Jaccard coefficients: 0.40 for conserved clusters and 0.50 for variable clusters, emphasizing the importance of using large sequence collections to achieve stable prediction. The transmission chains (Lemey et al., 2005; Leitner et al., 1996) contain few sequences compared with other datasets, and the overlap of the identified clusters for them is low: the overall Jaccard coefficient compared with clusters identified for the LANL dataset is 0.16, with Jaccard coefficients of 0.23 and 0.05 for clusters of extremely conserved and of extremely variable residues, respectively.

The pattern of extremely variable clusters is particularly noteworthy in cases when evolutionary pressure from the administered drugs may be involved. EuResist includes many sequences from
Table 3. Performance of different tools for predicting protein–protein interaction interfaces (average over all 10 HIV-1 proteins ± standard deviation)

<table>
<thead>
<tr>
<th>Tool</th>
<th>Accuracy (%)</th>
<th>PPV (%)</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our studyb</td>
<td>49.28 ± 11.03</td>
<td>62.62 ± 19.76</td>
<td>18.48 ± 10.00</td>
<td>82.75 ± 14.70</td>
</tr>
<tr>
<td>SPPIDER (Porollo and Meller, 2007)</td>
<td>60.80 ± 6.69</td>
<td>61.20 ± 11.61</td>
<td>64.59 ± 25.26</td>
<td>49.38 ± 33.57</td>
</tr>
<tr>
<td>PredUs (Zhang et al., 2010)</td>
<td>65.68 ± 14.32</td>
<td>67.20 ± 15.26</td>
<td>67.67 ± 20.75</td>
<td>61.75 ± 23.39</td>
</tr>
<tr>
<td>WHISCY (Vries et al., 2006)</td>
<td>55.86 ± 8.34</td>
<td>61.75 ± 21.40</td>
<td>60.33 ± 9.78</td>
<td>46.54 ± 10.73</td>
</tr>
<tr>
<td>PROMATE (Neuviirth et al., 2004)</td>
<td>55.65 ± 7.02</td>
<td>68.62 ± 9.33</td>
<td>41.70 ± 27.07</td>
<td>72.28 ± 26.05</td>
</tr>
</tbody>
</table>

For reverse transcriptase, capsid, nucleocapsid and Tat, PROMATE do not yield any predictions or predicts 100% of the surface to be a protein–protein interaction interface, thus these results were excluded from the comparison. Likewise, results of PredUs and SPPIDER for Tat and nucleocapsid, and of WHISCY for Tat were excluded. The best values are shown in bold italics.

bAccuracy = (true positives + true negatives)/all residues in the corresponding protein.

Our study = predicted extremely conserved and extremely variable clusters.

patients under therapy, and patients from the Belgian transmission chain (Lemey et al., 2005) (from which the protease had been sequenced) received various antiviral drugs. In agreement with that, the extremely variable clusters for EuResist dataset include the residue Ile93 (numbering according to HXB2 strain), a mutation of which is associated with resistance to atazanavir (Johnson et al., 2006). The extremely variable clusters in the protease for transmission chains comprise only two residues for β that corresponds to the largest silhouette value, none of which is associated with resistance to drugs. However, if we use the second-best silhouette value (which is only 0.08 less), the extremely variable clusters comprise five positions, of which Asp60 and Leu63 are associated with resistance to atazanavir and lopinavir, respectively (Johnson et al., 2006). Extremely conserved clusters in protease calculated using the LANL dataset are very small (only two residues for the largest silhouette value) and do not include any resistance-associated positions, probably owing to the large fraction of sequence from untreated patients in this set.

For HIV-2, SIV and FIV, we observe a good overlap of the identified clusters with those of HIV-1: the corresponding Jaccard coefficients for proteins of HIV-2 and SIV are 0.20, meaning that between a third and a half of the residues in either extremely conserved or extremely variable clusters in each virus are also found in such clusters in the other virus. For FIV, the overlap is much smaller, producing a Jaccard coefficient of only 0.09, in agreement with the larger evolutionary distance to FIV. The overlap is larger for extremely conserved clusters, except the ones identified for gp120 and integrase. For gp120, extremely conserved clusters in all immunodeficiency viruses other than HIV-1 are small, and they have no common position with gp120 from HIV-1. Extremely variable clusters, on the contrary, have at least one common position in all viruses, even as distant ones as FIV. This may point to some common antigenic properties of this protein across immunodeficiency viruses.

4 Discussion

4.1 Other methods for predicting protein–protein interaction interfaces

There exist a multitude of tools for predicting protein-protein interaction interfaces, which can be based on both sequence and structural features (Bradford and Westhead, 2005; Deng et al., 2009; Luo et al., 2014; Maleki et al., 2013; Srivastava et al., 2013; Wang et al., 2006). Sequence features can include residue conservation, sequence profiles and phylogeny. Structural features sometimes take into account solvent accessibility of residues and their physical and chemical properties. These methods are typically trained on all available structures of protein complexes in the Protein Data Bank. For some of them, a web interface for prediction is available (Neuviirth et al., 2004; Porollo and Meller, 2007; Vries et al., 2006), and we used these tools to predict protein–protein interaction interfaces for HIV-1 proteins. Then we compared the results in a manner similar with the one used for the clusters identified in this study: any residue located less than 5 Å away from any other protein chain in any experimentally resolved structure of an HIV-1 protein was considered to be a part of a protein–protein interaction interface, except those that were found to interact with antibodies.

On an average, conservation of protein–protein interaction interfaces does not significantly differ from the conservation of the whole protein surface or of all protein residues (Korkin et al., 2005). For HIV-1, conservation of the protein–protein interaction interfaces is 26±21%, of the surface 25±19%, and of all protein residues 26±21% (average and standard deviation calculated over all proteins). However, the extremely conserved clusters are enriched in interaction interfaces and correlate well with protein interaction hotspots (see Results). Compared with other tools, tight clustering is in fact capable of predicting the observed protein–protein interaction interfaces with very high specificity (Table 3).

It must be noted that on one hand our method is not specifically designed to distinguish protein–protein interaction interfaces from other types of interaction interfaces (i.e. with nucleic acids or low molecular-weight ligands), in this respect, it is fundamentally different from other tools for the prediction of protein–protein interaction interfaces. On the other hand, our method does not benefit from training data derived from the available three-dimensional structures of such interfaces, such as chemical properties that are typical to amino acid residues of these interfaces. This may be also regarded as an advantage of the presented approach, since it is capable of finding unusual interaction interfaces as well. The method was designed to identify small, but extremely conserved or extremely variable patches, which may account for relatively low values for accuracy and specificity.

We have also applied an alternative clustering approach to the LANL dataset. The approach is similar to the one described in Warren et al. (2013) with two modifications. First, instead of using exclusively 100% conserved residue positions, a floating threshold was introduced to account for a greater diversity of a much bigger dataset of HIV-1 sequences. Second, an advanced clustering algorithm, DBScan (Ester et al., 1996), was applied to obtain clusters of conserved residues. DBScan is a density-based clustering algorithm that also allows us to (1) filter the outliers and (2) rely exclusively on a distance matrix without explicitly defining the feature vectors. The results of this approach agree well with the extremely conserved clusters identified in this study:
clusters are presented in red; clusters from Warren et al. (2013) in green; overlap between the predictions in purple

Fig. 3. Clustering results for HIV-1 reverse transcriptase: extremely conserved clusters are presented in red; clusters from Warren et al. (2013) in green; overlap between the predictions in purple.

Fig. 4. Clusters of conserved residues for polymerase PA of Influenza A H1N1. Coloring scheme as in Figure 3.

4.2 Identification of conserved clusters in influenza a H1N1
We have additionally applied our tight clustering approach to identify extremely conserved and extremely variable clusters for proteins of the Influenza A H1N1 virus, which was originally studied in Warren et al. (2013), using the data from that study. The obtained conserved clusters agree well with the patches identified in Warren et al. (2013); the Jaccard index of the two sets is 0.31. For the Influenza A polymerase PA, both approaches show a strong overlap: 42% of residues detected by Warren et al. (2013) were also clustered with our tight clustering approach in the clusters of extremely conserved residues (Fig. 4). These residues also include the site involved in binding with the protein PB1 (Zürcher et al., 1996).

Acknowledgements
We are grateful to Nico Pfeifer and Thomas Lengauer for the critical reading of the manuscript.

Funding
This work was supported by the Department for Computational Biology and Applied Algorithmics, Max Planck Institute for Informatics, and National Science Foundation (DBI-0845196 to D.K.).

Conflict of Interest: none declared.

References
Maleki,M. et al. (2013) The role of electrostatic energy in prediction of obli-
McLellan,J. et al. (2011) Structure of HIV-1 gp120 V1/V2 domain with
Morrow,J. and Zhang,S. (2012) Computational prediction of hot spot resi-
Neuvirth,H. et al. (2004) ProMate: a structure based prediction program to identify

Shimokawa,H. et al. et al (2011) Structure of HIV-1 gp120 V1/V2 domain with
Morrow,J. and Zhang,S. (2012) Computational prediction of hot spot resi-
Neuvirth,H. et al. (2004) ProMate: a structure based prediction program to identify
O.S.Voitenko et al. (2011) Evolutionary insights into host–pathogen interactions.

Srivastava,B. et al. (2013) PPIcons: identification of protein–protein inter-
Thompson,J. et al. (1997) The CLUSTALX windows interface: flexible strate-
Thompson,J. et al. (1997) The CLUSTALX windows interface: flexible strate-

Trivedi,P. and Wang,N. (2014) Host immune responses accelerate pathogen
Diversity. ISME J., 8, 727–731.
Bioinformatics, 86, 324–345.
Wu,X. et al. (2000) Functional significance of conserved residues in the
Woolhouse,M. et al. (2002) Biological and biomedical implications of the co-
Wu,X. et al. (2011) Focused evolution of HIV-1 neutralizing antibodies re-
Zhang,Q. et al. (2010) Protein interface conservation across structure space.
Zhou,T. et al. (2010) Structural basis for broad and potent neutralization of