BMC Bioinformatics

BioMed Central

Research

Open Access

Pairwise and higher-order correlations among drug-resistance mutations in HIV-1 subtype B protease

Omar Haq[1], Ronald M Levy*[1,2], Alexandre V Morozov*[1,3] and Michael Andrec[1,2]

Address: [1]BioMaPS Institute for Quantitative Biology, Rutgers, the State University of New Jersey, 610 Taylor Road, Piscataway NJ 08854, USA, [2]Department of Chemistry and Chemical Biology, Rutgers, the State University of New Jersey, 610 Taylor Road, Piscataway NJ 08854, USA and [3]Department of Physics and Astronomy, Rutgers, the State University of New Jersey, 136 Frelinghuysen Road, Piscataway NJ 08854, USA

Email: Omar Haq - omarhaq@rci.rutgers.edu; Ronald M Levy* - ronlevy@lutece.rutgers.edu; Alexandre V Morozov* - morozov@physics.rutgers.edu; Michael Andrec - andrec@lutece.rutgers.edu

* Corresponding authors

from ECCB 2008 Workshop: Annotations, interpretation and management of mutations (AIMM)
Cagliari, Italy. 22 September 2008

Published: 27 August 2009

BMC Bioinformatics 2009, **10**(Suppl 8):S10 doi:10.1186/1471-2105-10-S8-S10

This article is available from: http://www.biomedcentral.com/1471-2105/10/S8/S10

© 2009 Haq et al; licensee BioMed Central Ltd.
This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: The reaction of HIV protease to inhibitor therapy is characterized by the emergence of complex mutational patterns which confer drug resistance. The response of HIV protease to drugs often involves both primary mutations that directly inhibit the action of the drug, and a host of accessory resistance mutations that may occur far from the active site but may contribute to restoring the fitness or stability of the enzyme. Here we develop a probabilistic approach based on connected information that allows us to study residue, pair level and higher-order correlations within the same framework.

Results: We apply our methodology to a database of approximately 13,000 sequences which have been annotated by the treatment history of the patients from which the samples were obtained. We show that including pair interactions is essential for agreement with the mutational data, since neglect of these interactions results in order-of-magnitude errors in the probabilities of the simultaneous occurence of many mutations. The magnitude of these pair correlations changes dramatically between sequences obtained from patients that were or were not exposed to drugs. Higher-order effects make a contribution of as much as 10% for residues taken three at a time, but increase to more than twice that for 10 to 15-residue groups. The sequence data is insufficient to determine the higher-order effects for larger groups. We find that higher-order interactions have a significant effect on the predicted frequencies of sequences with large numbers of mutations. While relatively rare, such sequences are more prevalent after multi-drug therapy. The relative importance of these higher-order interactions increases with the number of drugs the patient had been exposed to.

Conclusion: Correlations are critical for the understanding of mutation patterns in HIV protease. Pair interactions have substantial qualitative effects, while higher-order interactions are individually smaller but may have a collective effect. Together they lead to correlations which could have an

important impact on the dynamics of the evolution of cross-resistance, by allowing the virus to pass through otherwise unlikely mutational states. These findings also indicate that pairwise and possibly higher-order effects should be included in the models of protein evolution, instead of assuming that all residues mutate independently of one another.

Background

The protease enzyme coded for by the *pol* gene of the Human Immunodeficiency Virus HIV-1 plays a critical role in the reproduction of the virus by cleaving the GAG precursor protein in a sequence-specific manner into its functional form, and as such, is a key target of several families of commonly used drugs used to control HIV infection [1]. Unfortunately, the virus has been able to evolve resistance to many of these drugs, in part due to the high mutation rates in the HIV genome [2]. The patterns of mutations in protease are complex, involving multiple key primary mutations that inhibit the action of drugs and a host of accessory mutations that can modulate the enzyme's stability or activity or otherwise enhance the fitness of the virus. It is now understood that these mutations do not occur independently of each other, but instead are correlated, resulting in complex patterns of co-evolving mutations [3-7].

Previous studies have mostly focused on correlations between mutations in the HIV protease gene at the pair level [3,5-7]. However, recognition that the observed mutations may also be involved in higher-order interactions has led to a few studies in which correlated pairs of mutations are grouped using tools such as multidimensional scaling [3,6], Bayesian networks [8], networks defined by patterns of conditional selection pressure [5], and clustering [9,10]. The underlying assumption is that understanding higher-order interactions is important for a complete understanding of the evolution of resistance in HIV protease.

In this paper, we investigate correlations among HIV protease mutations at and beyond the pair level, and the impact of drug treatment on the nature of those correlations. We only consider the presence or absence of a non-synonymous mutation relative to a defined wild-type sequence, and not the precise base or amino acid substitution which has occurred. We develop a hierarchy of probabilistic log-linear models [11] that can in principle describe residue interactions of arbitrary order, and use those to analyze HIV protease sequence data obtained from patient cohorts with varying protease inhibitor (PI) treatment histories.

We use "connected information" [12] to quantify inter-residue interactions at the triplet and higher level. Unlike the Bayesian network approach [13], the information-theoretic methodology allows us to distinguish intrinsic three-body effects from the cases in which correlations between three random variables can be attributed mostly to pairwise interactions. The connected information viewpoint of higher order correlation has not been previously used in the analysis of mutational patterns in HIV protease, although it has been employed in a much more limited analysis of the V3 loop of the HIV envelope protein [14], and log-linear models have been used to study protein-protein interactions [15]. We find that pairwise interactions are necessary to achieve even qualitative agreement with the mutational data, while higher order interactions play an important role in predicting how frequently sequences with several mutations appear in the database. Simultaneous appearance of multiple mutations may play an important role in the phenomenon of multiple- or cross-resistance of the viral protease.

Results

Increased mutation frequencies under drug exposure

As has been previously observed [3], we find that the overall number of mutations seen in HIV protease increases significantly with the number of PIs that the patient has been exposed to. This is seen in Figure 1, where we show the distribution of the number of mutations at drug-associated positions for sequences isolated from the drug-

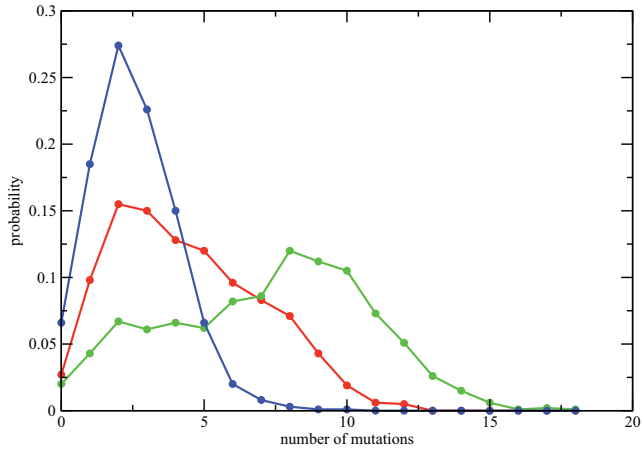

Figure 1
Probability distributions for the total number of mutations among the 41 drug-associated residues in HIV protease for sequences obtained from the PI0 (blue), PI1 (red), and PI2+ (green) cohorts.

naive cohort (PI0), from a PI monotherapy cohort (PI1), and from a cohort treated with 2 or more PIs (PI2+), as estimated from the database described in the Methods. For some residue positions, the increase in the mutation frequency between the PI0 and PI2+ cohorts is nearly eightfold, while other positions show no discernible change (Additional File 1). This observed increase in mutation frequencies at drug-associated sites is largely responsible for the shift in the distribution shown in Figure 1.

It is of interest to ask which amino acid positions exhibit elevated mutation frequencies under drug treatment. Mutations at many of these positions are associated with decreased HIV-1 inhibitor susceptibility, and it is useful to classify mutations as belonging to "primary" vs "accessory" resistance classes. The terms "secondary" and "compensatory" have also been used as synonyms for "accessory". The specific criteria for such a classification are *ad hoc* in nature, but have generally been defined as follows.

Primary mutations are usually selected first in the presence of the drug and confer resistance, even when present as single point mutations [16,17]. They can be structurally important, e.g. situated near the enzymatic active site, in which case their effect on inhibitor binding can be rationalized due to their physical proximity to the inhibitor [18]. In the case of protease, however, there are exceptions, as some mutations (such as positions 54, 76, 88 and 90) are situated far from the active site or have no direct contacts with the substrate, yet still reduce drug susceptibility [16]. The mechanism of action of these mutations is not clearly understood [19].

Accessory mutations confer resistance only when present with additional primary or accessory drug-resistance mutations and have little or no effect on inhibitor susceptibility on their own. Some of these mutations occur in the absence of drug treatment, but their frequency of occurrence is observed to increase in treated patients. Accessory mutations may "rescue" possible losses of activity or stability in the enzyme that may have been caused by a destabilizing primary resistance mutation, and therefore may have a compensatory function in restoring viral fitness [20].

We make use of a primary and accessory classification scheme based on the work of Shafer et al. [21-23,17]. We define 17 primary drug resistance positions (residues 23, 24, 30, 32, 33, 46, 47, 48, 50, 53, 54, 73, 76, 82, 84, 88, 90) and 24 accessory drug resistance positions (residues 10, 11, 13, 20, 34, 35, 36, 43, 45, 55, 58, 60, 63, 71, 74, 75, 77, 79, 83, 85, 89, 91, 93, 95). The remaining positions are polymorphic mutations not associated with drug resistance or are conserved sites [24,3,25,21,26,23].

As seen in Figure 2, the positions which exhibit the most elevated mutation frequencies in PI2+ relative to PI0 sequences are for the most part primary and accessory drug resistance positions. It is interesting to note that sequences bearing mutations at any one of the 6 residues with the most elevated mutation frequencies (10, 46, 54, 71, 82 and 90) are resistant to most of the current PIs, including amprenavir (APV), indinavir (IDV), nelfinavir (NFV), ritonavir (RTV), saquinavir (SQV), atazanavir (ATV), and lopinavir (LPV). For instance, according to the Stanford HIV database [16], the common mutations at these residues (L10I, M46I, I54V, A71V, V82A and L90M) are associated with six or all seven of the drugs mentioned above. On the other hand, primary drug resistance positions that show the least amount of change in their mutation frequency (23, 30, 76 and 88) are generally inhibitor-specific. For example, D30N is associated only with NFV treatment, L23I and L76V are specific to two drugs each (NFV and SQV, and IDV and LPV, respectively), while N88D is specific to IDV, SQV and NFV treatment. Thus, positions that provide drug resistance to multiple inhibitors are more frequently mutated in the PI2+ cohort than positions that are specific to a small number of protease inhibitors. Upon examination of 278 ligand-bound crystal structures of HIV-1 protease, no heavy atom of residues 10, 46, 54, 71, or 90 is ever within 3.4 Å of any ligand bound at the cleft (data not shown). The atoms of residue

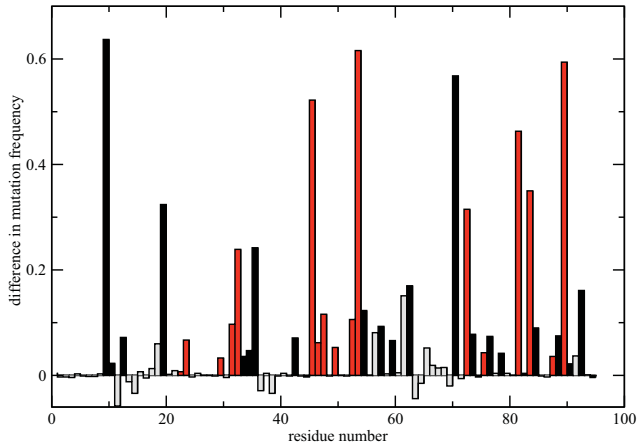

Figure 2
Difference in the frequency of mutated residues between patients treated with 5PIs and PI-naive patients for all positions in HIV protease. Red bars correspond to primary drug resistnce positions, black bars are accessory drug resistance positions, and grey bars are positions not associated with drug resistance.

82, however, do contact ligands in 54% of the crystal structures examined.

Exhaustive analysis of residue pairs

We begin by investigating pair correlations and their association with the structure of HIV protease. Several groups have studied pair correlations as a means of identifying functionally dependent residues in the HIV protease and other systems [6,5,4,3,7]. For each of the $\binom{82}{2}$ = 3, 321 pairs of positions, a 2 × 2 contingency table was constructed for both the drug-naive sequences and the sequences treated by two or more drugs. The binomial or "product moment" [11] correlation coefficient

$$\phi_{AB} = \frac{P(A_m B_m) - P(A_m)P(B_m)}{\sqrt{P(A_m)P(A_0)P(B_m)P(B_0)}} \quad (1)$$

was calculated for both datasets, where for amino acid position A we denote the wild-type state as A_0 and the mutated state as A_m. Of the 3, 321 φ values, 98 from the drug-naive and 223 from the treated set were considered to be statistically significant with substantial correlations ($|\varphi| > 0.1$, $p < 0.001$). These correlations match qualitatively with a prior study which used the same database but had fewer sequences: the φ values of the top 15 positively correlated pairs for the PI2+ cohort and those of Wu, et al. [3] have a Spearman rank order correlation coefficient of 0.80 (data not shown). Furthermore, we observe larger φ values for pairs of drug-associated positions compared to non-drug associated pairs (Figure 3), which is consistent with previous observations [3].

Previous studies of pair correlations in protein families have indicated that coevolving pairs of residues tend to lie closer to each other in structure than random residue pairs [27-31]. In the case of the HIV protease, the distribution of φ values for drug-associated positions with distance shows a characteristic triangular shape, particularly for the PI2+ cohort (Figure 3B) [32]. In particular, distances associated with the most correlated residue pairs (30–88, 54–82, 32–47, and 37–77) are all within a few Ångstroms of each other. The triangular shape in Figure 3B is not surprising, given that most of the drug-associated positions are on the substrate cleft and thus tend to be relatively close to each other in space. As a result, 78 pairs (36% of the statistically significant pairs) are within 8 Å of each other. We also see no tendency for exposed residues to preferentially coevolve, in contrast to previous studies on mutation covariation in other protein families [27,29,15].

Of the 26 statistically significant negatively correlated pairs in the PI2+ cohort, 10 pairs involve either residues 30 or 88. Residue 30 is negatively correlated with positions 82, 10, 46, 90, 54, 73 and 84 (in decreasing order of the magnitude of the correlation), while position 88 is negatively correlated with positions 82, 73 and 54. It is interesting to note that position 63, which has a high mutation rate in both the PI0 and PI2+ cohorts, is negatively correlated with positions 80, 52, 5, 83, 64, and 61, of which only position 83 is associated with drug resistance [3]. In fact, of the 23 unique positions involved in the 26 negatively correlated pairs, most, but not all, are positions of drug resistance. It is possible that the 7 non-drug associated positions, 5, 15, 20, 52, 61, 64 and 80 play a role in the stability or function of the protein, even if they do not interfere with inhibitors [33-35]. Residue 80, in particular is negatively correlated with three residues, 63, 71 and 90, all of which play either primary or accessory roles in drug resistance.

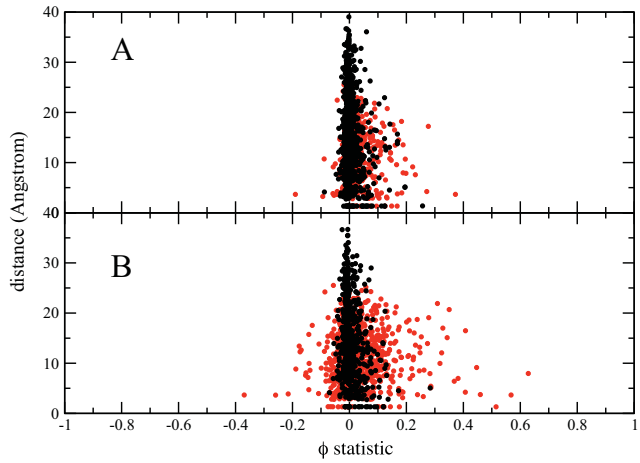

Figure 3
Scatterplot of distance vs φ statistic for drug associated positions in red and non-drug associated positions in black for the PI0 (A) and PI2+ (B) cohorts. The value along the y-axis is the closest distance between any two heavy atoms of the two residues based on the crystal structure of wild type protease (PDB ID 1PRO).

Drug treatment has a significant impact on the pair correlations, as can be seen in Figure 4. Of the $\binom{82}{2}$ = 3, 321 pairs of positions, only 100 have statistically significant ($p < 0.001$) φ values and are common to both the PI0 and PI2+ cohorts, and of these, most of the positively correlated pairs in the PI0 cohort become more strongly correlated in the PI2+ cohort. However, some pairs of residues which are weakly positively correlated in the PI0 cohort become negatively correlated in the PI2+ cohort. There are 12 such pairs, almost all of which have at least one primary drug resistance position, and 8 of them involve

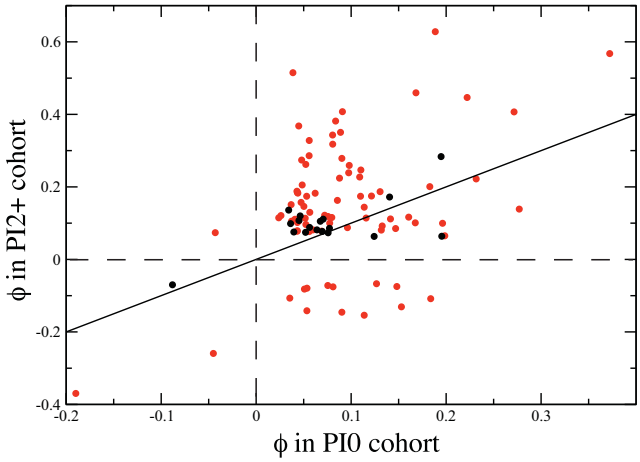

Figure 4
Scatterplot showing the change in φ statistic upon drug treatment, with drug associated positions in red and non-drug associated positions in black. Only pairs with statistically significant (p < 0.001) pair correlations are shown. The solid line corresponds to no change upon drug treatment.

either residue 30 or 88. It is interesting to note that in PI0 cohort sequences, residue 30 is positively correlated with positions 24, 46, 54, 84 and 90, but becomes strongly negatively correlated in the PI2+ cohort. The anticorrelation of residue 30 with the other primary positions after drug treatment has been previously observed experimentally [36] and in a prior statistical study [3], but it is not clear why this anticorrelation exists only in the presence of drugs.

Additionally, the types of residues involved in pair correlations changes upon treatment, with accessory positions becoming more prominent: 10% of the correlated pairs in the PI0 cohort consist of a primary and an accessory position, and that this combination increases to 31% in the PI2+ cohort. Furthermore, 45% of the pairs in the PI0 cohort involve at least one accessory position, which increases to 61% in the PI2+ cohort. As expected, pairs of non-resistance-associated positions decrease from 52% to 35%. It is interesting to note that the same trend is not observed for primary positions: 57% of pairs in the PI0 cohort contain at least one primary position, and this is essentially unchanged in the PI2+ cohort (59%). Therefore, drug treatment causes correlated pairs involving primary and non-resistance associated positions to be replaced by pairs involving primary and accessory positions.

Exhaustive analysis of residue triples

To study interactions among mutations beyond the pair level, we quantify the amount of information in the observed distribution that cannot be explained by pair correlation. This is done using the three-body "connected information" [12], which is defined as the difference in Shannon entropy $S(P) = -\sum_i p_i \log p_i$ between the distributions $P(A, B, C)$ and $\tilde{P}^{(2)}(A, B, C)$, where the latter is the maximum entropy distribution subject to the constraints that all of its univariate and bivariate marginals are the same as that of $P(A, B, C)$:

$$I_c^{(3)}(A,B,C) = S[\tilde{P}^{(2)}(A,B,C)] - S[P(A,B,C)]. \quad (2)$$

Similarly, we define the two-body connected information to be

$$I_c^{(2)}(A,B,C) = S[P(A)P(B)P(C)] - S[\tilde{P}^{(2)}(A,B,C)]. \quad (3)$$

The total information arising from correlation at any level is given by the "multi-information" or the Kullback-Leibler divergence between the observed distribution and the prediction based on an independent model:

$$I_{multi}(A,B,C) = \sum_{abc} P(a,b,c) \log \frac{P(a,b,c)}{P(a)P(b)P(c)} = S(A) + S(B) + S(C) - S(A,B,C). \quad (4)$$

The maximum entropy distribution $\tilde{P}^{(2)}(A, B, C)$ can be thought of as being "in between" the independent model and the observed distribution, since it is more constrained than the independent model but does not have the full correlation structure of the observed data (Figure 5). Since $I_c^{(2)} + I_c^{(3)} = I_{multi}$, we can interpret the two- and three-body connected information as the part of the correlation that can be explained by pair interactions alone, and that which arises from three-body interactions, respectively [12].

We examined the degree of connected information in both the PI0 and PI2+ cohort sequences. The connected information $I_c^{(3)}$ was calculated for all $\binom{82}{3} = 88,560$ residue triples in the HIV protease as described in the Methods section below. In the PI0 cohort sequences, only 175 residue triples have statistically significant three-body interaction at the $p = 0.001$ level, while the PI2+ cohort sequences have 6,300 significant triples. Furthermore, the significant triples from the PI2+ cohort are enriched in drug resistance associated positions: 32% of them consist

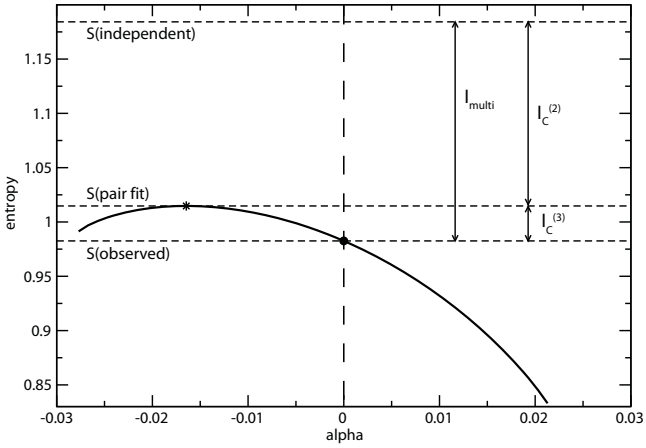

Figure 5

The Shannon entropy for the family of all triplet distributions having the same univariate and bivariate marginals as the data for the triplet 46-54-90 (PI2+ cohort) plotted as a function of the parameter α defined in Equation 8. The observed distribution is indicated by the filled circle, and the entropies of the independent model (S_{ind}), the model with only pair interactions (S_{pair}) and the observed data (S_{obs}) are indicated, as well as I_{multi} and the connected information measures $I_C^{(2)}$ and $I_C^{(3)}$. S_{pair} is the entropy of the distribution denoted by $\tilde{P}^{(2)}$ (A, B, C) in the text which has the maximum entropy subject to the constraints that all of the univariate and bivariate marginals are the same as the observed distribution, and is denoted in the figure by a star.

of only drug associated positions, compared to 12% of the full set of 88,560. In contrast, the significant triples from the PI0 cohort show no such enrichment: only 9% consist solely of drug resistance associated positions.

The 10 triples with the largest $I_c^{(3)}$ values from the PI2+ cohort sequences are shown in Table 1, along with the corresponding values of I_{multi}. Residues from the largest $I_c^{(3)}$ triple (46-54-90) are also displayed on the structure of the HIV protease (Figure 6). As can be seen in Table 1, the total contribution of three-body interactions to the information content of the observed data is at most 10%. Although this is a relatively small effect, there is a clear association with drug therapy, with the largest $I_c^{(3)}$ values increasing substantially with the number of PIs the patient was exposed to (Figure 7). There is relatively little overlap between the largest 100 $I_c^{(3)}$ triples in the PI0 and PI2+ cohorts, with only 11 triples in common. The impact

Table 1: The 10 triples in HIV protease from the PI2+ cohort sequences with the largest $I_C^{(3)}$ (*A*, *B*, *C*) values (all three-body interactions are significant with p <<> 10^{-6}).

residue triple	$I_C^{(3)}$ (*A*, *B*, *C*)	I_{multi} (*A*, *B*, *C*)
46-54-90	0.03219	0.20170
46-71-90	0.03101	0.29041
46-82-90	0.03085	0.20640
82-84-90	0.02570	0.13237
10-46-90	0.02445	0.31475
46-71-73	0.02195	0.16610
36-46-90	0.02191	0.15949
46-54-71	0.02072	0.23595
46-77-82	0.02068	0.14573
20-82-90	0.02030	0.14661

of these higher-order interactions is particularly manifest in the probability of the occurrence of three simultaneous mutations in a given residue triple (Figure 8), which is significantly increased compared to what would be predicted based on a pair model for sequences from the PI2+ cohort (but not in sequences from the PI0 cohort).

While all of the triples in Table 1 have three-body interactions that are highly statistically significant, it is also important to obtain a practical feeling for the magnitude of these interactions. Let us consider the 46-54-90 triple. In Figure 9 we show the correlation between the predicted and observed probabilities for each of the 8 binary states for the independent model (red) and the two-body fit (black) for this triple in the PI2+ cohort. It is clear that the independent model fits the data poorly: the probability that all three residues are mutated is underestimated by an order of magnitude, and some of the others are over- or underestimated by factors of 2 or 3. Although the two-body fit does a better job of reproducing the observed probabilities, it still leads to noticeable deviations from the straight line for the 46-54-90 triple.

Another way in which the magnitude of the effect of the three-body interactions can be visualized is by considering conditional probabilities for the 46-54-90: $P(54_m|46_m90_m)$ = 0.4186, $P(54_m|46_m90_0)$ = 0.5508, $P(54_m|46_090_m)$ = 0.4359, and $P(54_m|46_090_0)$ = 0.0280. All of these probabilities differ significantly from the independent estimate of $P(54_m)$ = 0.1178. We also see clear evidence of "triplet correlation" in the data, in the sense that the probability of a mutation depends very strongly on the state of both of the other residues, e.g. residue 54 is much less likely to be mutated if *both* 46 and 90 are wild-type than if only one of them is wild-type.

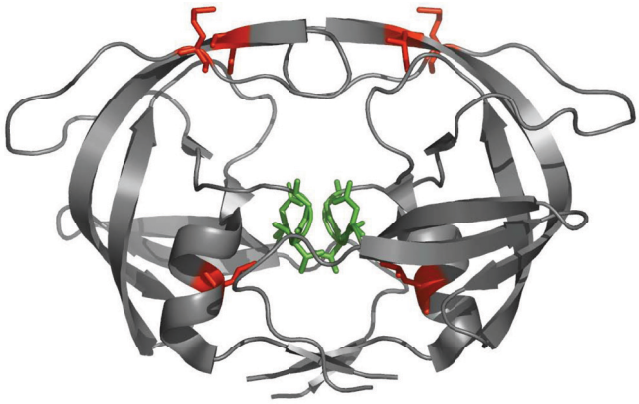

Figure 6

Structure of the HIV protease dimer (PDB code 1HN0). The catalytic triad (residues 25, 26, and 27) is highlighted in green, while the residues from the triple with the largest $I_C^{(3)}$ in the PI2+ cohort (46-54-90, see Figure 9) are shown in red. Residue 90 is close to the catalytic triad, whereas residues 46 and 54 are close to each other, with the smallest distance between heavy atoms of 3.11 Å).

Much of this probabilistic dependency, however, can be accounted for by pair interactions. The corresponding probabilities for the best-fit two-body model are $\tilde{P}^{(2)}(54_m|46_m90_m) = 0.6286$, $\tilde{P}^{(2)}(54_m|46_m90_0) = 0.2236$, $\tilde{P}^{(2)}(54_m|46_090_m) = 0.2316$, and $\tilde{P}^{(2)}(54_m|46_090_0) = 0.0488$. Even though this model contains no three-body interactions, a qualitative "triplet correlation" of the type seen above remains (i.e. $\tilde{P}^{(2)}(54_m|46_m90_0) \approx \tilde{P}^{(2)}(54_m|46_090_m) > \tilde{P}^{(2)}(54_m|46_090_0)$), indicating that non-trivial three-way probabilistic dependencies can arise purely from pair correlations [37]. Overall, three-body interactions do quantitatively modulate the probabilities, but only to a small degree, since $I_c^{(3)}$ is an order of magnitude smaller than I_{multi} even for the triples in Table 1. In other words, the contribution of three-body interactions to any of the triplet distributions that describe the mutational patterns of protease taken three at a time is roughly 10% or less of the effect induced by the pairwise interactions acting on these positions. Nonetheless, there can be substantial effects on a "micro level", such as the more than twofold difference in $P(54_m|46_m90_0)$ seen here.

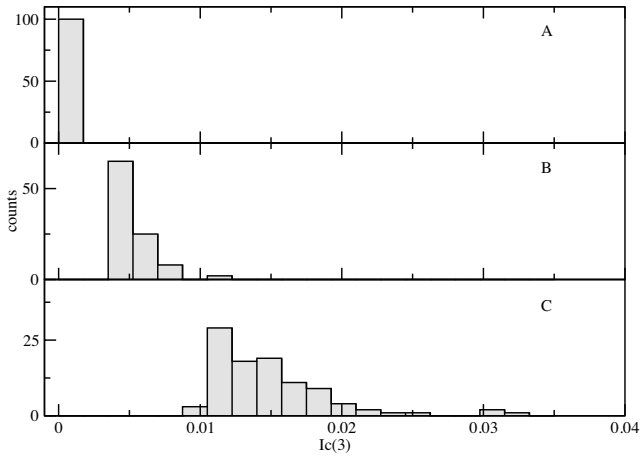

Figure 7

Distribution of the 100 largest $I_C^{(3)}$ values for the PI0 (A), PI1 (B), and PI2+ (C) cohorts.

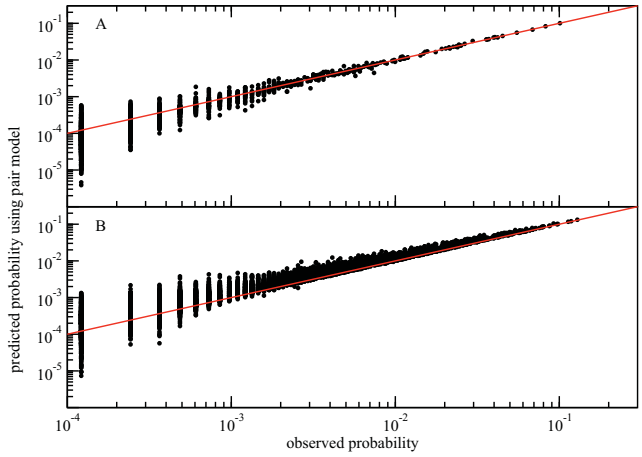

Figure 8

Predicted vs observed probabilities for a triple mutant for all triples of drug-associated positions in the HIV protease using sequences from the PI0 (A) and PI2+ (B) cohorts. Mutational states that were unobserved in the database and would have an observed probability maximum likelihood estimate of zero are not shown. The dots correspond to the best fit pair-term model, and the red lines of slope 1 correspond to the perfect agreement of the predicted probabilities with those observed, which would be obtained if the three-body interactions were included.

Increased higher-order interactions in larger residue groups

The small but consistent systematic deviations seen in Figure 8B raises the possibility that these interactions could combine synergystically to produce more substantial effects over larger clusters of residues. Ideally, this would be studied by fitting log-linear models to increasingly larger clusters using the data from the PI2+ cohort. How-

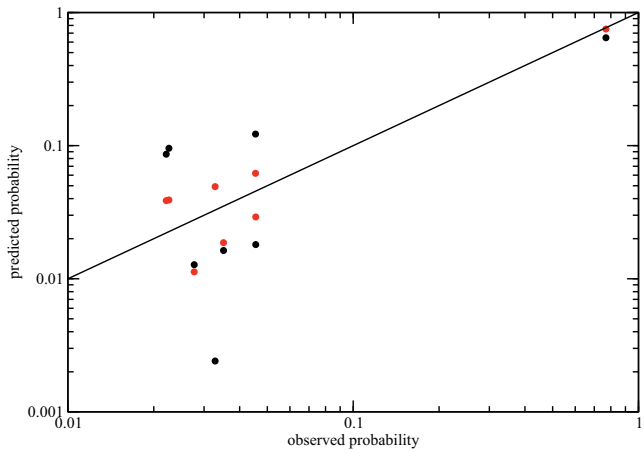

Figure 9

Predicted vs observed probabilities for the residue triple in HIV protease with the largest $I_C^{(3)}$ in the PI2+ cohort (46-54-90). The black dots correspond to the best-fit independent model, while the red dots correspond to the best-fit pair-term model (Equation 11). The solid line of slope 1 corresponds to perfect agreement of the predicted data with the observed, which would be obtained if the three-body term were included.

ever, the size of cohort limits our ability to do this to clusters of no more than ≈ 15 residues. We begin by examining the 10-residue group 20-32-46-48-53-54-58-74-82-90, which was chosen to have the largest higher-order interactions from among a limited set of residues defined by the three key primary drug resistance positions 30, 82 and 90, and the known accessory positions associated with them. For this 10-residue group in the PI2+ cohort, we observe strong pair interactions which bring observed and predicted probabilities into qualitative, order-of-magnitude agreement, weaker three-body interactions which further improve the agreement, and very weak four-body and higher interactions which have quantitative impact on a small number of state probabilities (Additional File 2). This can be quantified in terms of connected information: from the entropies of the of the observed, three-body model, pair model, and independent distributions (Figure 10), we find that $I_C^{(2)} = S(ind) - S(pair) = 0.4785$, $I_C^{(3)} = S(pair) - S(trip) = 0.1187$, and the sum of the remaining connected information measures of fourth order and higher is $S(trip) - S(obs) = 0.0703$ (these information theoretic measures were found to be robust with respect to sampling error as determined by bootstrap). When compared to $I_{multi} = S(ind) - S(obs) = 0.6676$, we see that three-body and higher-order interactions make up 28% of the total correlation information. This is substantial increase over the (at most) 10% contribution from higher-order interactions to the observed triplet distributions. By contrast, the same 10-residue group for the PI0 cohort displays a substantially smaller overall degree of correlation, as seen by the small I_{multi} in Figure 10.

We also studied the distribution of the total number of mutations in the same 10-residue group (20-32-46-48-53-54-58-74-82-90). Appropriate subsets of the state probabilities determined above were summed to obtain the distribution of the total number of mutated residues for the independent, two-body, and three-body models, and the results are shown in Figure 11A. The distribution for the independent model is very different from the observed distribution: the probability of having no mutations is considerably underestimated, and the upper tail is much too thin. Adding pair terms greatly improves the "no mutation" probability and considerably extends the length of the tail. The tail length is further modulated by the addition of the three-body interactions, bringing the distribution very close to the observed probabilities.

It should be noted that since the univariate marginals are preserved by all of the models, the mean total number of mutations is the same for all 4 curves in Figure 11A. Therefore, under- or overestimation of the total number of mutations in one part of the distribution must be compensated by over- or under-estimation (respectively) in

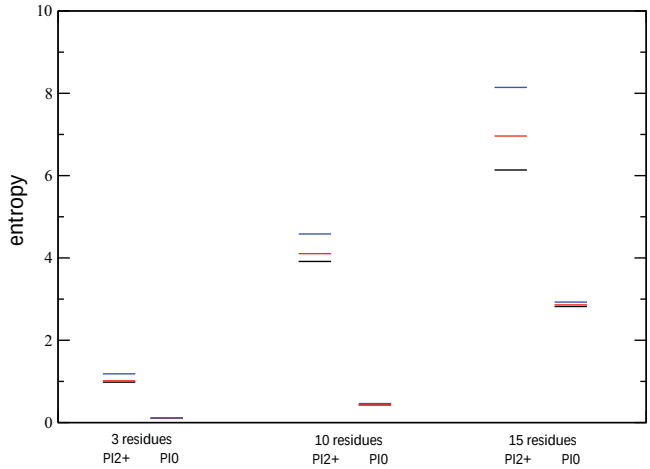

Figure 10

Shannon theoretic entropy for the observed (black), best pair-term model (red) and independent model (blue) for the three-residue group 46-54-90, the 10-residue group 20-32-46-48-53-54-58-74-82-90, and the 15-residue group 10-20-33-36-46-54-55-63-71-73-74-82-84-90-93 for both the PI0 and PI2+ cohorts.

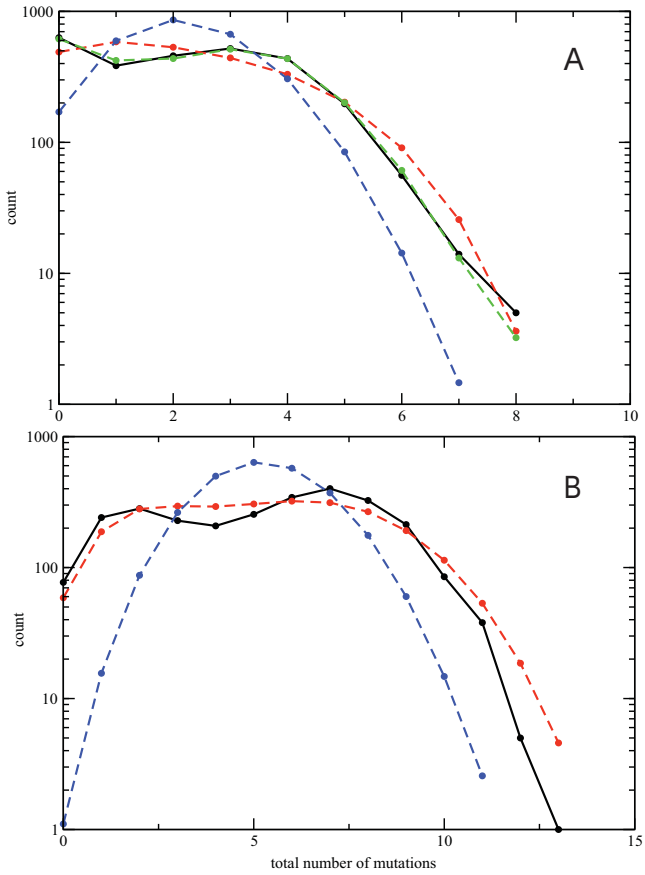

Figure 11
Total number of observed (solid black curve) or predicted mutations in the 2,702 sequences of the PI2+ cohort for the 10-residue group 20-32-46-48-53-54-58-74-82-90 (A) and the 15-residue group 10-20-33-36-46-54-55-63-71-73-74-82-84-90-93 (B). The dashed curves represent the predictions for the independent (blue), pair-term (red), and pair+three-body model (green, not fit for the 15-residue group).

another, implying that the curves must cross one other. To quantify this effect, we can compare the predicted probabilities for seeing 5 or more mutations under each model and comparing to the observed probability. Those probabilities are 0.0372, 0.1197, and 0.1030 for the independent, pair-term model, and pair+three-body model, respectively, compared to the observed probability of 0.1007. The deviations of the first two from the observed are highly statistically significant, while the latter has a p-value of ≈ 0.06. The qualitative distribution is well accounted for by the pair interaction model, with the three-body and higher interactions modulating the details of the shape of the distribution, such as the increase in frequency of 3–5 and 13–15 mutations, at the expense of a decrease in the frequency of 6–10 mutations.

To see if the synergistic effects seen for the 10-residue group become even stronger for 15 residues, we repeated this analysis for the 15-residue group 10-20-33-36-46-54-55-63-71-73-74-82-84-90–93 (Figure 10), which was chosen by selecting the residues with the largest change in mutation frequencies upon PI treatment. For the PI2+ cohort, we find that $I_C^{(2)} = S(ind) - S(pair) = 1.1810$, the sum of the remaining connected information measures of third order and higher is $S(pair) - S(obs) = 0.8249$ and $I_{multi} = S(ind) - S(obs) = 2.0059$. The contribution of three-body and higher-order interactions now make up 41% of the total correlation information. However, this result may somewhat overestimate the true amount of higher-order correlation. A fit with a three-body model (Equation 12) gives an estimate of $I_C^{(3)}$ of 0.2638, and the ratio $I_C^{(3)}/(I_C^{(2)} + I_C^{(3)}) = 18\%$ represents a lower bound on the contribution from correlations beyond the pair level. A more complete account of the many issues involved in estimating the amount of higher order correlation and its precision and accuracy for finite data sets will be the subject of a future communication. The comparison of the predicted and observed distributions of total number of mutations bears out this result (Figure 11B), showing more pronounced differences between the observed (black) and pair-model (red) distributions. In addition to the overestimation of the upper tail similar to that seen for the 10-residue group, we now also see that the pair model cannot reproduce the bimodal shape seen in the observed data. Again, the same 15-residue group for the PI0 cohort shows considerably weaker overall correlation (Figure 10).

Discussion and conclusion

Treating HIV protease with drugs results in the appearance of complex mutational patterns: observed mutations are not limited to the active site and often occur in groups that involve two or more residues. Furthermore, some mutations occur even in the absence of drugs, presumably following neutral rather than adaptive evolution. To study correlations between different residue positions in HIV protease, we have developed a hierarchy of models that allows us to include inter-residue correlations of arbitrary order within a consistent framework. Using only HIV protease sequences as input, we find that pair interactions become common and quite strong after PI treatment. In fact, it is often impossible to achieve even qualitative agreement with the data without including the two-body terms (Figures 9, 11, and Additional File 2). This finding calls into question a common assumption employed in

current probabilistic approaches to phylogeny [38] that most residues evolve independently.

We have developed an information-theoretic method to study interactions between mutations beyond the pairwise level. Our approach is based on the notion of the connected information $I_c^{(3)}$ (Equation 2) [12]. While there are a variety of quantitatively different measures of pair correlation [11,39] that may differ in their sensitivity in various regimes, they all measure essentially the same qualitative feature of the observed data. On the other hand, no single summary statistic can capture all of the various characteristics of higher-order behavior, leading to multiple descriptions that provide complimentary information. Connected information is one intuitive statistic that provides insight into the degree of structure in the data beyond the pair correlation level.

Connected information provides information which is complementary to Bayesian network analyses based on factorizations of the joint probability. It can readily verified that $I_C^{(3)} = 0$ if at least one of the random variables is independent of the other two, or if the joint distribution involves conditional independence (e.g. $P(A, B, C) = P(A)P(B|A)P(C|A)$). However, a joint distribution with triplet-level probabilistic dependencies (in the sense that two of the variables are independent of each other, but the third depends jointly on the state of the other two, e.g. $P(A, B, C) = P(A)P(B)P(C|A, B)$, or if $P(A, B, C)$ cannot be factorized into any simpler form) could still be consistent with no three-body connected information if the observed triplet distribution is the maximum entropy distribution relative to its marginals. Thus, even if a Bayesian network-style analysis shows that a given triple cannot be factorized into any simpler form, that "triplet correlation" could still be consistent with a very small or zero $I_c^{(3)}$, indicating that the observed behavior is dominated by two-body interactions. In fact, it has been shown that very complex correlation patterns among random variables can arise from large numbers of weak pairwise interactions [37].

Other information-theoretic measures of "higher-order correlation" have also been proposed, including higher-order mutual information, which measures "frustration" or the degree of synergy vs redundancy among several random variables [40]. While this measure has been used in the analysis of HIV envelope protein sequence data [41], its interpretation is considerably less intuitive. Similarly, *ad hoc* methods for finding putative clusters of mutually correlated residues [3,9,8,10] cannot reliably uncover sets that have intrinsic higher order interactions, as defined by large $I_c^{(3)}$ (data not shown).

Plotting φ values for pairs of residues as a function of the distance between them (Figure 3B) reveals that while some large pair correlations arise from direct contacts between residues (e.g. $\varphi \approx 0.8$, $d < 5$ Å), there are also strong correlations ($\varphi \approx 0.5$) between amino acids separated by 15 Å or more, making physical coupling between them very unlikely. To provide an example of the former, we consider mutations involving residues 30 and 88. The closest distance between heavy atoms of residues 30 and 88 is just 3.66 Å, making likely some sort of physical interaction between them. Mutations at residue 30 are strongly and uniquely associated with resistance to the protease inhibitor nelfinavir, and there exists a strong correlation between mutations at positions 30 and 88 [42] which may be due in part to a compensation of the loss of a surface negative charge from the D30N mutation being restored by N88D [9].

It is possible that chains of intermediate interactions result in long-range coupling between two coevolving yet physically distant residues [43]. However, because non-zero values of λ_{ij} indicate a presence of *direct* interactions between residues *i* and *j* in our model [15], we can decompose such "energetically connected pathways" into contributions from separate pairs. In contrast to a previous study [15], we find that non-zero values of λ_{ij} are only weakly correlated with distance (Additional File 3). This lack of correlation is not entirely surprising, since even direct interactions between a pair of residues need not have a purely physical origin. Indeed, if protein fitness is a non-linear function of its stability or enzymatic activity [44], two mutations can be correlated because they compensate each other by making independent and opposite contributions to the overall fitness, even if there is no direct or indirect physical interaction between them [45]. By the same argument, the three-body terms also result from a mixture of physical and epistatic (compensatory) origins.

We have shown that three-body and higher-order correlations have the largest effect on the probabilities of the simultaneous occurrence of multiple mutations in the HIV protease (Figure 11). Since both this and previous studies have found that the total number of mutated positions is correlated with treatment by multiple protease inhibitors (Figure 1) [3], the presence of higher-order interactions may influence how protease reacts to multiple drugs, and could have an important impact on the evolution of cross-resistance, for example, by providing

the virus with an "escape hatch" of large numbers of mutations. Higher-order interactions could also impact the time evolution of mutations by allowing the virus to pass through otherwise unlikely mutational states. We have seen that the impact of higher-order interactions in 10 to 15 residue clusters is at least a factor of two larger than the largest $I_C^{(3)}$ values for residue triples (approximately 20% or more of the total entropy change). One of the outstanding questions raised by this work is whether the impact of higher-order interactions for HIV evolving under the pressure of multiple drugs continues to become stronger for larger residue groups (ultimately the set of all 41 drug-associated positions). Unfortunately, there is not enough sequence data to perform such an analysis. Short of obtaining additional data, it may also be possible to explore this question by constructing synthetic data sets using λ_i and λ_{ij} values consistent with an observed $I_C^{(3)}$ distribution at the level of residue triples (i.e. Figure 7C).

The sequence-based approach presented here is not limited to the HIV protease and its response to drug treatment, and should be equally useful in studies of the evolution of drug resistance in other systems. Moreover, it will be of interest to extend our techniques to other examples of short-term neutral and adaptive evolution, including controlled evolution in the lab accompanied by protein sequencing at different timepoints. Recent work has suggested that evolutionary pathways of proteins are relatively restricted and may be predictable in general [46], and specific methods for predicting the mutational dynamics of HIV protease have been proposed, based on Bayesian network models [47] or pairwise conditional selection pressure [5]. A better understanding of the nature of the probabilistic dependencies underlying the network models should lead to improved prediction strategies. However, our model cannot distinguish between physical and epistatic origins of the observed co-evolution. To do this, we need a different approach which would explicitly introduce protein fitness as a function of residue energies (including interactions across protein-protein and protein-ligand interface). These energies would be fit against the sequence data, resulting in a prediction that decomposes observed inter-residue correlations into the physical and epistatic parts. This approach is currently being pursued in our laboratories.

Methods

HIV sequence database

Aligned and annotated HIV-1 protease amino acid sequences were obtained using the web interface of the Stanford HIV Drug Resistance Database http://hivdb.stanford.edu/cgi-bin/PI_Form.cgi[16]. The sequences are all classified under the HIV-1 Main group, subtype B. For the purposes of this paper, we are considering only the number of protease inhibitors and not the specific combinations of protease inhibitors used in the treatment. There are 13,608 sequences in this curated set, of which 8,229 sequences are drug naive, 2,677 sequence are associated with PI monotherapy, and the remaining 2,702 sequences are associated with between 2 through 6 protease inhibitors.

Calculation of higher-order interactions

For residue triples, we compute $I_c^{(3)}$ by writing the triplet probability distribution in log-linear form [11]:

$$P(A,B,C) = \exp(\lambda_0 + \lambda_A A + \lambda_B B + \lambda_C C + \lambda_{AB} AB + \lambda_{BC} BC + \lambda_{AC} AC + \lambda_{ABC} ABC), \tag{5}$$

where we assign numerical values to the states of A, B and C, e.g. 0 and 1. We find $\tilde{P}^{(2)}(A, B, C)$ by setting $\lambda_{ABC} = 0$ in Equation 5 and fitting the six parameters λ_i and λ_{ij} to the values that maximize the likelihood of the data under a multinomial model [48]. However, in the triplet case it is possible to avoid direct nonlinear optimization: let us represent the 8 observed probabilities by the vector

$$\mathbf{p}_0 \equiv (p_{000}, p_{00m}, p_{0m0}, p_{0mm}, p_{m00}, p_{m0m}, p_{mm0}, p_{mmm}), \tag{6}$$

where, e.g. $p_{0m0} = P(A_0, B_m, C_0)$. It is sufficient to consider only the three marginals $P(A_m)$, $P(B_m)$, and $P(C_m)$, and three suitably chosen bivariate marginals, e.g. $P(A_m, B_m)$, $P(A_m, C_m)$, and $P(B_m, C_m)$, since the remaining 9 bivariate marginals can be reconstructed as combinations of these: $P(A_m, B_0) = P(A_m) - P(A_m, B_m)$, $P(A_0, B_m) = P(B_m) - P(A_m, B_m)$, $P(A_0, B_0) = 1 - P(A_m) - P(B_m) + P(A_m, B_m)$, etc. The six marginals can then be written as a matrix equation involving $\mathbf{p}_0$:

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix} \mathbf{p}_0^T = \begin{pmatrix} P(A_m) \\ P(B_m) \\ P(C_m) \\ P(A_m, B_m) \\ P(A_m, C_m) \\ P(B_m, C_m) \end{pmatrix}. \tag{7}$$

Since the matrix multiplying $\mathbf{p}_0$ is rectangular with dimensions 6 × 8, it has a two-dimensional null space, with basis vectors $\mathbf{n}_1 = (1, 0, 0, 0, 0, 0, 0, 0)$ and $\mathbf{n}_2 = (0, -1, -1, 1, -1, 1, 1, -1)$. Then, any linear combination $\alpha_1\mathbf{n}_1 + \alpha_2\mathbf{n}_2$ added to $\mathbf{p}_0$ will not change the marginals. However, α_1 and α_2 cannot be chosen independently without violating the normalization of $\mathbf{p}_0$: we must choose $\alpha_1 = \alpha_2 = \alpha$. There-

fore, the family of all possible distributions that have the same univariate and bivariate marginals are mapped out by the parameter α using the relation

$$\mathbf{p}_\alpha = \mathbf{p}_0 + \alpha(1,-1,-1,1,-1,1,1,-1), \quad (8)$$

where the feasible values of α are constrained by the non-negativity requirement for probabilities. Furthermore, $\lambda_{ABC} = 0$ in Equation 5 implies that

$$\frac{p_{000}p_{mm0}p_{0mm}p_{m0m}}{p_{0m0}p_{m00}p_{00m}p_{mmm}} = 1. \quad (9)$$

Therefore, to find $\tilde{P}^{(2)}(A, B, C)$ it suffices to find the value of α which satisfies Equation 9:

$$\frac{(p_{000}+\alpha)(p_{mm0}+\alpha)(p_{0mm}+\alpha)(p_{m0m}+\alpha)}{(p_{0m0}-\alpha)(p_{m00}-\alpha)(p_{00m}-\alpha)(p_{mmm}-\alpha)} = 1, \quad (10)$$

leading to a cubic equation in α [11].

To obtain the maximum entropy distributions for more than three binary random variables, nonlinear optimization is unavoidable. However, instead of directly maximizing the entropy subject to the marginal probability constraints, we maximize the likelihood subject to the constraints that the λ's vanish beyond a given order [48]. In most cases, the latter will be more computationally efficient since the number of λ variables grows polynomially with the number of variables, while the dimensionality of the null space defined by the marginal probability constraints (which is one-dimensional for three variables) increases exponentially with the number of variables.

In general, we fit data on mutation and wild-type amino acid counts to the following hierarchy of probabilistic models: the independent model $P(A, B, C,...) = P(A)P(B)P(C)\cup$, the "two-body" model:

$$P(A,B,C,...) = \exp(\lambda_0 + \sum_i \lambda_i I + \sum_{ij} \lambda_{ij} IJ), \quad (11)$$

and the "three-body" model:

$$P(A,B,C,...) = \exp(\lambda_0 + \sum_i \lambda_i I + \sum_{ij} \lambda_{ij} IJ + \sum_{ijk} \lambda_{ijk} IJK), \quad (12)$$

where λ is the vector of parameters, the indices (i, j) and (i, j, k) run over all distinct combinations of $\{A, B, C,...\}$ with $i \neq j$ and $i \neq j \neq k$, respectively, and I, J, and K are numerical values of the corresponding random variable (we use 0 for wild-type and 1 for mutant). For an n-variate distribution $P(A, B, C,...)$ there are $n(n - 1)/2$ pair parameters λ_{ij} and $n(n - 1)(n - 2)/6$ three-body parameters λ_{ijk} (λ_0 is a normalization constant). The independent model was determined by forming products of the observed univariate marginals. The magnitudes of the λ_i and λ_{ij} parameters in the two-body model are related to the mutation frequencies at site i and pair correlations between sites i and j, respectively. In fact, the magnitudes of λ_{ij} in the context of a two-body model have been proposed as a measure of "direct information", i.e. the part of pair correlation resulting from direct coupling [15]. It should be noted that the relative magnitudes of λ_{ij} is dependent on the choice of the numerical values assigned to the random variables I, J and K in Equations 11 and 12. In this work, we assign values of 0 and 1 for wild-type and mutant, respectively, for computational convenience. It has been argued, however, that a more appropriate choice of numerical values is one which is symmetric about zero, e.g. ± 1, which allows "gauge constraints" to be introduced [15]. While this choice will affect the the values of λ_{ij} and their interpretation as "direct information", it will not change the best-fit two-body or three-body probabilities and consequently will have no impact on the values of $I_c^{(3)}$ and related measures of higher-order interactions.

The two-body model for $n = 3$ was fit by solving Equation 10 exactly [49]. If no feasible solution of Equation 10 exists, then $I_c^{(3)}$ was set to zero. For $n \geq 4$, the unknown parameters in Equations 11 and 12 were determined by maximizing the multinomial log-likelihood

$$L(\lambda) = \sum_i N_i \ln P(i), \quad (13)$$

where i is one of the 2^n states, N_i is the number of times that state was observed, and $P(i|\lambda)$ is the predicted probability for state i to be observed given the vector of parameters λ. Maximization was performed numerically using the "nlm" function of the R software package [50]. All entropy and connected information values are given in natural log units. Statistical significance of the three-body interactions was estimated using the likelihood ratio test under the null hypothesis that the data were generated by $\tilde{P}^{(2)}(A, B, C)$ by Monte Carlo sampling. For all of the residue triples in Table 1, the p-values for the observed likelihood ratio were too small to be estimated (p <<> 10^{-6}),

indicating very strong statistical significance for the three-body interaction.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

OH and MA wrote software and performed the analyses, and OH, RML, AVM and MA wrote the paper.

Additional material

Additional File 1

Mutation frequencies for selected positions in the HIV protease as a function of the number of PIs the patient was exposed to. *Data is shown for residues 10 (black, solid line), 54 (red, solid line), 90 (green), 71 (blue), 46 (orange), 77 (black, dashed line), and 35 (red, dashed line).*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-S8-S10-S1.pdf]

Additional File 2

Predicted vs observed probabilities for the 2_{10} mutational states of the ten-residue group 20-32-46-48-53-54-58-74-82-90 in HIV protease for the PI2+ cohort. *Mutational states that were unobserved in the database and would have an observed probability maximum likelihood estimate of zero are not shown. The black dots correspond to the best-fit independent model, the red dots correspond to the best-fit pair-term model (Equation 11), and the green dots correspond to the best-fit three-body model (Equation 12). The solid line of slope 1 corresponds to perfect agreement of the predicted data with the observed, which would be obtained if all higher-order terms were included.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-S8-S10-S2.pdf]

Additional File 3

Scatterplot of distance vs λ_{ij} parameters estimated using* I *and* J *values of ± 1 as described in the Methods for the 15-residue group 10-20-33-36-46-54-55-63-71-73-74-82-84-90-93 in the PI2+ cohort. *The value along the* y*-axis is the closest distance between any two heavy atoms of the two residues based on the crystal structure of wild type protease (PDB ID 1PRO).*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-S8-S10-S3.pdf]

Acknowledgements

This work has been supported by a grant from the National Institutes of Health, GM-30580 (to RML) and by an Alfred P Sloan Research Fellowship (to AVM).

This article has been published as part of *BMC Bioinformatics* Volume 10 Supplement 8, 2009: Proceedings of the European Conference on Computational Biology (ECCB) 2008 Workshop: Annotation, interpretation and management of mutations. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcbioinformatics/10?issue=S8.

References

1. Swanstrom R, Erona J: **Human immunodeficiency virus type-1 protease inhibitors: Thereputic successes and failures, suppression and resistance.** *Pharmacol Ther* 2000, **86:**145-170.
2. Rambaut A, Posada D, Crandall KA, Holmes EC: **The causes and consequences of HIV evolution.** *Nat Rev Genet* 2004, **5:**52-61.
3. Wu TD, Schiffer CA, Gonzales MJ, Taylor J, Kantor R, Chou S, Israelski D, Zolopa AR, Fessel WJ, Shafer RW: **Mutation patterns and structural correlates in human immunodeficiency virus type 1 protease following different protease inhibitor treatments.** *J Virol* 2003, **77:**4836-4847.
4. Hoffman NG, Schiffer CA, Swanstrom R: **Covariation of amino acid positions in HIV-1 protease.** *Virology* 2003, **314:**536-548.
5. Chen L, Lee C: **Distinguishing HIV-1 drug resistance, accessory, and viral fitness mutations using conditional selection pressure analysis of treated versus untreated patient samples.** *Biology Direct* 2006, **1:**14.
6. Rhee SY, Liu TF, Holms SP, Shafer RW: **HIV-1 subtype B protease and reverse transcriptase amino acid covariation.** *PLoS Comput Biol* 2007, **3:**e87.
7. Wang Q, Lee C: **Distinguishing functional amino acid covariation from background linkage disequilibrium in HIV protease and reverse transcriptase.** *PLoS ONE* 2007, **2:**e814.
8. Deforche K, Silander T, Camacho R, Grossman Z, Soares MA, Van Laethem K, Kantor R, Moreau Y, Vandamme A-M: **Analysis of the HIV-1 *pol* sequences using Bayesian Networks: implications for drug resistance.** *Bioinformatics* 2006, **22:**2975-2979.
9. Svicher V, Ceccherini-Silberstein F, Erba F, Santoro M, Gori C, Bellocchi MC, Gianella S, Trotta MP, Monforte A, Antinori A, Perno CF: **Novel human immunodeficiency virus type 1 protease mutations potentially involved in resistance to protease inhibitors.** *Antimicrob Agents Chemother* 2005, **49:**2015-2025.
10. Liu Y, Eyal E, Bahar I: **Analysis of correlated mutations in HIV-1 protease using spectral clustering.** *Bioinformatics* 2008, **24:**1243-1250.
11. Bishop YM, Fienberg SE, Holland PW: *Discrete Multivariate Analysis: Theory and Practice The MIT Press*; 1975.
12. Schneidman E, Still S, Berry MJ II, Bialek W: **Network information and connected correlations.** *Phys Rev Lett* 2003, **91:**238701.
13. Pearl J: *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference Morgan Kaufmann Publishers*; 1988.
14. Ahn C, Koch GG, Paynter L, Preisser JS, Seillier-Moiseiwitsch F: **Log-linear models for mutations in the HIV genome.** *J Stat Plan Inf* 2007, **137:**3227-3239.
15. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T: **Identification of direct residue contacts in protein-protein interaction by message passing.** *Proc Natl Acad Sci USA* 2009, **106:**67-72.
16. Kantor R, Betts BJ, Ravela J, Rhee SY, Gonzales MJ, Shafer RW: **Human immunodeficiency virus reverse transcriptase and protease sequence database.** *Nucleic Acids Res* 2003, **31:**298-303.
17. Johnson VA, Brun-Vézinet F, Clotet B, Günthard HF, Kuritzkes DR, Pillay D, Schapiro JM, Richman DD: **Update of the drug resistance mutations in HIV-1: December 2008.** *Topics in HIV Medicine* 2008, **16:**138-145.
18. Clemente JC, Hemrajani R, Blum LE, Goodenow MM, Dunn BM: **Secondary mutations M36I and A71V in the human immunodeficiency virus type 1 protease can provide an advantage of the emergence of the primary mutation D30N.** *Biochemistry* 2003, **42:**15029-15035.
19. Muzammil S, Ross P, Freire E: **A major role for a set of non-active site mutations in the development of HIV-1 protease drug resistance.** *Biochemistry* 2003, **42:**631-638.
20. Shafer RW, Dupnik K, Winters MA, Eshleman SH: **A guide to HIV-1 reverse transcriptase and protease sequencing for drug resistance studies.** In *Human retroviruses and AIDS, theoretical biology and biophysics Los Alamos, NM: Los Alamos National Laboratories*; 2001:1-51.
21. Rhee SY, Taylor J, Wadhera G, Ben-Hur A, Brutlag DL, Shafer RW: **Genotypic predictors of human immunodeficiency virus type 1 drug resistance.** *Proc Natl Acad Sci USA* 2006, **103:**17355-17360.
22. Shafer RW, Schapiro JM: **HIV-1 drug resistance mutations: An updated framework for the second decade of HAART.** *AIDS Rev* 2008, **10:**67-84.
23. Shafer RW, Rhee SY, Pillay D, Miller V, Sandstrom P, Schapiro JM, Kuritzkes DR, Bennett D: **HIV-1 protease and reverse tran-**

scriptase mutations for drug resistance surveillance. *AIDS* 2007, **21:**215-223.
24. Hertogs K, Bloor S, Kemp SD, Van de Eynde C Van den, Alcorn TM, Pauwels R, Van Houtte M, Staszewski S, Miller V, Larder BA: **Phenotypic and genotypic analysis of clinical HIV-1 isolates reveals extensive protease inhibitor cross-resistance: A survey of over 6000 samples.** *AIDS* 2000, **14:**1203-1210.
25. Ohtaka H, Schon A, Freire E: **Multidrug resistance to HIV-1 protease inhibition requires cooperative coupling between distal mutations.** *Biochemistry* 2003, **42:**13659-13666.
26. Johnson VA A, Brun-Vézinet F, Clotet B, Günthard HF F, Kuritzkes DR R, Pillay D, Schapiro JM M, Richman DD D: **Update of the drug resistance mutations in HIV-1 2007.** *Topics in HIV Medicine* 2007, **15:**119-125.
27. Pollock DD, Taylor WR, Goldman N: **Coevolving proteing residues: Maximum likelihood indentification and relationship to structure.** *Journal of Molecular Biology* 1999, **287:**187-198.
28. Tiller ERM, Lui TWH: **Using muliple interdependency to separate functional from phylogenetic correlations in protein alignments.** *Bioinformatics* 2003, **19:**750-755.
29. Gloor GB, Martin LC, Wahl LM, Dunn SM: **Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions.** *Biochemistry* 2005, **44:**7156-7165.
30. Flares MA, Travers SAA: **A novel method for detecting intramolecular co-evolution: Adding a further dimension to selective constraints analyses.** *Genetics* 2006, **173:**9-23.
31. Perez-Jimenez R, Godoy-Ruiz R, Parody-Morreale A, Ibarra-Molero B, Sanchez-Ruiz JM: **A simple tool to explore the distance distribution of correlated mutations in proteins.** *Biophys Chem* 2006, **119:**240-246.
32. Fodor AA, Aldrich RW: **Influence of conservation on calculations of amino acid covariance in multiple sequence alignments.** *Proteins* 2004, **56:**211-221.
33. Nijhuis M, Schuurman R, de Jong D, Erickson J, Gustchina E, Albert J, Schipper P, Gulnik S, Boucher CAB: **Increased fitness of drug resistant HIV-1 protease as a result of acquisition of compensatory mutations during suboptimal therapy.** *AIDS* 1999, **13:**2349-2359.
34. Perno CF, Cozzi-Lepri A, Balotta C, Forbici F, Violin M, Bertoli A, Facchi G, Pezzotti P, Cadeo G, Tositti G, Pasquinucci S, Pauluzzi S, Scalzini A, Salassa B, Vincenti A, Phillips A, Dianzani F, Appice A, Angarano G, Monno L, Ippolito G, Moroni M, Monforte A, Italian Cohort Naive Antiretroviral (I.CO.N.A.) Study Groupa: **Secondary mutations in the protease region of human immunodeficiency virus and virologic failure in drug-naive patients treated with protease inhibitor-based therapy.** *The Journal of Infectious Diseases* 2001, **184:**983-991.
35. Piana S, Carloni P, Rothlisberger U: **Drug resistance in HIV-1 protease: Flexibility-assisted mechanism of compensatory mutations.** *Protein Science* 2002, **11:**2393-2402.
36. Sugiura W, Matsuda Z, Yokomaku Y, Hertogs K, Larder B, Oishi T, Okano A, Shiino T, Tatsumi M, Matsuda M, *et al.*: **Interference between D30N and L90M in selection and development of protease inhibitor-resistant human immunodeficiency virus type 1.** *Antimicrob Agents Chemother* 2002, **46(3):**708-715.
37. Schneidman E, Berry MJ II, Segev R, Bialek W: **Weak pairwise correlations imply strongly correlated network states in a neural population.** *Nature* 2006, **440:**1007-1012.
38. Durbin R, Eddy S, Krogh A, Mitchison G: *Biological Sequence Analysis Cambridge University Press*; 1998.
39. Gower JC: *Measures of Similarity, Dissimilarity, Distance, Encyclopedia of Statistical Sciences* 1985, **5:**397-405.
40. Matsuda H: **Physical nature of higher-order mutual information: Intrinsic correlations and frustration.** *Phys Rev E* 2000, **62:**3096-3102.
41. Gilbert PB, Novitsky V, Essex M: **Covariability of selected amino acid positions for HIV type 1 subtypes C and B. AIDS Res.** *Human Retrovir* 2005, **21:**1016-1030.
42. Patick AK, Duran M, Cao Y, Shugarts D, Keller MR, Mazabel E, Knowles M, Chapman S, Kuritzkes DR, Markowitz M: **Genotypic and phenotypic characterization of human immunodeficiency virus type 1 variants isolated from patients treated with the protease inhibitor nelfinavir.** *Antimicrob Agents Chemother* 1998, **42:**2637-2644.
43. Steve W: **Lockless and Rama Ranganathan. Evolutionarily conserved pathways of energetic connectivity in protein families.** *Science* 1999, **286:**295-299.
44. Mustonen V, Kinney J Jr, Callan CG, Lässig M: **Energy-dependent fitness: A quantitative model for the evolution of yeast transcription factor binding sites.** *Proc Natl Acad Sci USA* 2008, **105:**12376-12381.
45. DePristo MA, Weinreich DM, Hartl DL: **Missense meanderings in sequence space: A biophysical view of protein evolution.** *Nat Rev Genet* 2005, **6:**678-687.
46. Weinreich DM, Delaney NF, DePristo MA, Hartl DL: **Darwinian evolution can follow only very few mutational paths to fitter proteins.** *Science* 2006, **312:**111-114.
47. Deforche K, Camacho R, Van Laethem K, Lemey P, Rambaut A, Moreau Y, Vandamme A-M: **Estimation of an *in vivo* fitness landscape experienced by HIV-1 under drug selective pressure useful for prediction of drug resistance evolution during treatment.** *Bioinformatics* 2008, **24:**34-41.
48. Good IJ: **Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables.** *Ann Math Stat* 1963, **34:**911-934.
49. Press WH, Teukolsky SA, Vetterling WT, Flannery BP: *Numerical Recipes in C: The Art of Scientific Computing Cambridge University Press, Cambridge*; 1992.
50. R Development Core Team: *R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria*; 2005.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

BioMedcentral