# Computational Mutagenesis Studies of Protein Structure-Function Correlations

**Majid Masso, Zhibin Lu, and Iosif I. Vaisman***
*Laboratory for Structural Bioinformatics, School of Computational Sciences, George Mason University, Manassas, Virginia*

*ABSTRACT* **Topological scores, measures of sequence-structure compatibility, are calculated for all 1,881 single point mutants of the human immunodeficiency virus (HIV)-1 protease using a four-body statistical potential function based on Delaunay tessellation of protein structure. Comparison of the mutant topological score data with experimental data from alanine scan studies specifically on the dimer interface residues supports previous findings that 1) L97 and F99 contribute greatly to the Gibbs energy of HIV-1 protease dimerization, 2) Q2 and T4 contribute the least toward the Gibbs energy, and 3) C-terminal residues are more sensitive to mutations than those at the N-terminus. For a more comprehensive treatment of the relationship between protease structure and function, mutant topological scores are compared with the activity levels for a set of 536 experimentally synthesized protease mutants, and a significant correlation is observed. Finally, this structure-function correlation is similarly identified by examining model systems consisting of 2,015 single point mutants of bacteriophage T4 lysozyme as well as 366 single point mutants of HIV-1 reverse transcriptase and is hypothesized to be a property generally applicable to all proteins. Proteins 2006; 64:234–245. © 2006 Wiley-Liss, Inc.**

**Key words: protein structure-function correlation; statistical geometry; computational mutagenesis; HIV-1 protease; HIV-1 reverse transcriptase; T4 lysozyme**

## INTRODUCTION

Unlike eukaryotic aspartyl proteases, retroviral proteases are only functional as a homodimer. Human immunodeficiency virus (HIV)-1 protease is responsible for cleavage of the gag and gag-pol polyprotein precursors, an essential step in the assembly of mature, infectious virus particles. Areas of particular interest among the 99 amino acids comprising the HIV-1 PR monomer [Fig. 1(A)] include the dimer interface (residues P1-T4 and C95-F99 from each monomer, forming an interdigitated, antiparallel N-C′-C-N′ β-sheet), the active site triad (D25-T26-G27), and the flexible flap region (M46-V56).[1,2]

Owing to the high error rate of HIV-1 reverse transcriptase (RT), mutations are readily introduced into the primary sequence of HIV-1 protease. Whereas substitutions occurring at evolutionarily conserved positions are rarely if ever tolerated, other positions in protease are capable of withstanding certain types of mutations. The latter include mutations that, while maintaining various degrees of enzyme activity, greatly reduce the binding affinity for protease inhibitors. An effective approach for gauging the relative contributions of constituent amino acid residues on the overall stability, adaptability, flexibility, and ligand binding capability of HIV-1 protease is via systematic mutational studies.[3–8]

Experimental studies of mutations in proteins are expensive and time-consuming, and thus the number of mutants included in such studies is limited. However, computational mutagenesis investigations can be performed on a large number of mutants, and with efficient algorithms and software, all possible mutants of a given protein can be analyzed. Several computational mutagenesis studies have been described recently in the literature that rely on readily available sequence and structure information, including physical and chemical properties of the substituted residues.[9–14]

Using a computational geometry technique based on Delaunay tessellation of protein structure, we previously reported on results of a comprehensive mutational analysis of HIV-1 protease.[15] First, a topological score measuring overall sequence-structure compatibility of protease was calculated by application of a Delaunay tessellation-derived four-body statistical potential function. Next, individual residue environment scores were also evaluated for each of the 99 positions in the protease monomer and plotted in a three-dimensional–one-dimensional (3D-1D) profile. Finally, a comprehensive mutational profile (CMP) was obtained for protease, whereby a score was calculated at each residue position that captures the mean environmental change caused by all possible residue substitutions. The 3D-1D and CMP profiles for protease were strongly inversely correlated ($R^2 = 0.88$).

By incorporating additional physical and evolutionary features of individual residues in the analysis, further important insights into structure-function correlations in

Grant sponsor: PhRMA Faculty Development Award in Bioinformatics (1998–2000).

*Correspondence to: Iosif I. Vaisman, Laboratory for Structural Bioinformatics, School of Computational Sciences, George Mason University, Manassas, VA 20110. E-mail: ivaisman@gmu.edu

Zhibin Lu's present address is Microarray Centre, 200 Elizabeth Street, Toronto, Ontario M5G 2C4, Canada.

Received 25 August 2005; Revised 30 December 2005; Accepted 6 January 2006

Published online 14 April 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20968

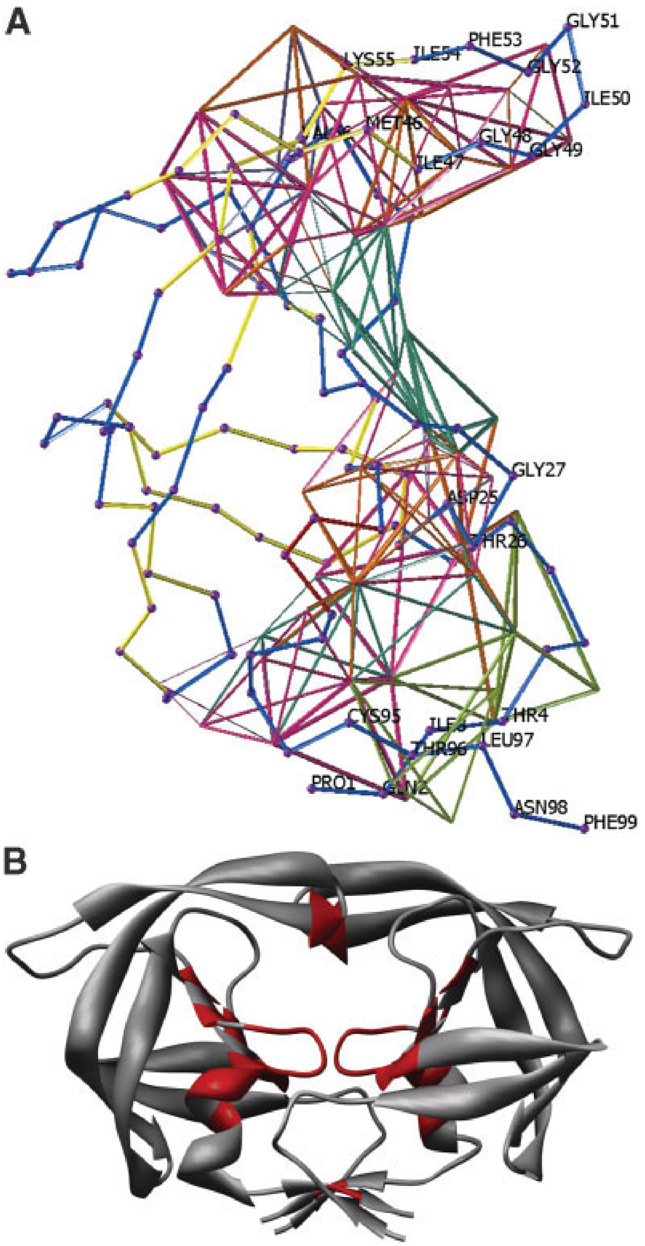


Fig. 1. **A:** Delaunay tessellation of a monomer of HIV-1 protease, highlighting the dimer interface (P1-T4 and C95-F99), the active site (D25-G27), and the flap region (M46-V56) residues along the backbone of the protein chain. The vertices used for the Delaunay simplices are the center of mass coordinates of the constituent amino acid side-chains. Only simplices whose edge lengths are all less than 10.4 Å are shown. **B:** Ribbon diagram of a functional homodimer of HIV-1 protease, highlighting the locations of 16 conserved residue positions derived from a multiple sequence alignment of 17 retroviral proteases (Fig. 4). The ribbon diagram was produced using the UCSF Chimera package.[32]

HIV-1 protease may be derived. In the current study, we begin by identifying evolutionarily conserved residue positions in protease and by correlating the level of conservation of residue positions with the level of sensitivity to mutations. Next, we analyze dimer interface mutants of protease in the context of the Delaunay tessellation approach, by comparing the topological score data with published experimental results from alanine scans performed at these residue positions. In the final step, we investigate the connection between structure and function in HIV-1 protease as follows. Every single point mutant of protease has an associated *residual* sequence-structure compatibility score, defined as the difference between the mutant and wild-type (WT) topological scores and initially obtained for the purpose of calculating the CMP. Theoretically there are a total of 99 residues × 19 substitutions/residue = 1,881 single point HIV-1 protease mutants. The residual scores of 536 of these mutants are compared with published data from a study in which the activity levels of these experimentally synthesized single missense protease mutants, located throughout the primary sequence of the protein, were measured relative to WT.[16] Although limited by the extent of the experimental data available (28% of all possible single point mutants), the investigation reveals a striking structure-function correlation.

We also perform a similar study on two additional systems: T4 lysozyme and HIV-1 RT. Residual scores are compared with mutant activity data obtained from published experiments involving 2,015 bacteriophage T4 lysozyme single point mutants[17] as well as 366 HIV-1 RT single point mutants.[18] A significant correlation between residual scores of the mutants and their activity levels is demonstrated for both of these model systems as well. The results lead us to conclude that our topological score data encode the relevant information about mutant structural changes from WT that is necessary to accurately reflect the influence of protein structure on function.

## MATERIALS AND METHODS

### Tessellation and Scoring Function

Delaunay tessellation of a protein structure, represented by the coordinates of the $C_\alpha$ atoms of each of the constituent amino acid residues, yields an aggregate of non-overlapping, space-filling, irregular tetrahedra.[19–21] The vertices of each Delaunay simplex in a protein tessellation objectively define a set of four nearest-neighbor residues in 3D space. The Quickhull algorithm is used to perform the protein tessellations.[22] A suite of programs has been written to perform all of the data extraction and formatting prior and subsequent to tessellation, as well as all of the calculations and data analyses subsequent to tessellation as described below.

In general, assuming that residue composition of Delaunay simplices is order independent, the 20 naturally occurring amino acids are capable of yielding 8,855 distinct quadruplets as vertices for a simplex.[20,21] For each quadruplet, a log-likelihood score is defined by $q_{ijkl} = \log(f_{ijkl}/p_{ijkl})$, where $f_{ijkl}$ represents the frequency of the quadruplet compositions containing residues $i$, $j$, $k$, $l$ in a non-redundant training set of high-resolution structures with low primary sequence identity obtained from the Protein Data Bank (PDB),[23] and $p_{ijkl}$ is the frequency of random occurrence of the quadruplet.[20,21] These 8,855 quadruplets of amino acid residues, along with their respective log-likelihood scores, define the four-body statistical potential function.

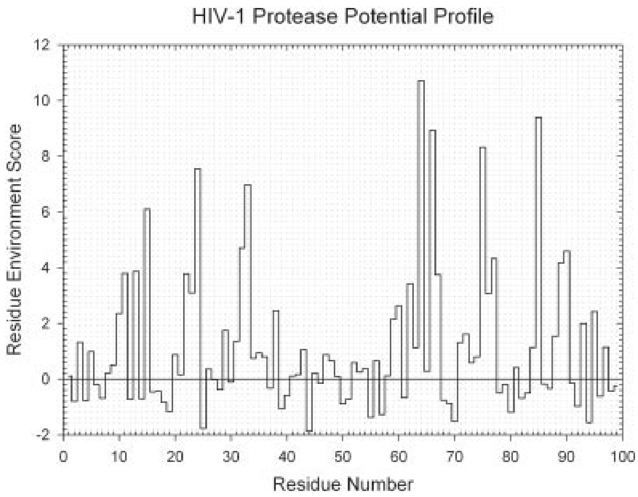


Fig. 2. A 3D-1D profile of a wild-type HIV-1 protease monomer (PDB ID: 3phv). Highest scoring residues tend to be found in the hydrophobic core, whereas lower scoring residues are exposed. A local minimum occurs at the active site catalytic residue D25. The topological score of the protease is 27.93.

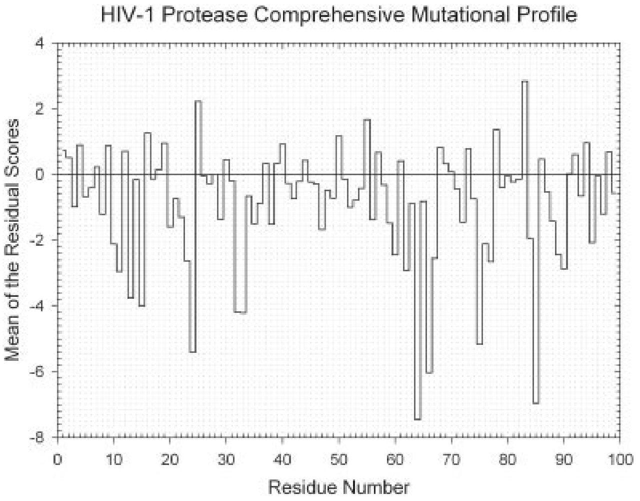


Fig. 3. CMP of an HIV-1 protease monomer. At each residue position, the plot reflects the mean of the residuals (the difference between mutant and wild-type topological scores) associated with all possible amino acid substitutions. The CMP is strongly inversely correlated with the 3D-1D profile in Figure 2 ($R^2 = 0.88$).

Using the above scoring function, a log-likelihood score can be assigned to each Delaunay simplex in a sample protein tessellation, based on the identity of the residue quadruplet whose $C_\alpha$ coordinates form the vertices of the simplex. The topological score of the protein is obtained by summing the log-likelihood scores of all the simplices forming the Delaunay tessellation of the protein structure.[20,21] A residue environment score can also be calculated for each amino acid position by summing only the log-likelihood scores of simplices that use the $C_\alpha$ coordinate of the residue position as a vertex. A graphical display of all the individual residue scores in a protein is referred to as a 3D-1D profile (Fig. 2). The PDB codes of the protein structures analyzed for this article include 3phv (HIV-1 protease monomer), 1g35 (HIV-1 protease dimer), 3lzm (bacteriophage T4 lysozyme), and 1rtjA (HIV-1 RT).

## CMP

The CMP is based on the tessellation of the WT protein. All 19 possible substitutions at each residue position are individually evaluated by changing only the identity of the WT amino acid while maintaining the original $C_\alpha$ coordinate. A mutation at a particular residue position impacts the log-likelihood scores of all the simplices that use the corresponding $C_\alpha$ coordinate as a vertex. Hence, the individual residue scores of all amino acids with $C_\alpha$ coordinates participating in these simplices are affected, as is the overall topological score of the protein. For each mutant protein, a *residual* is calculated as the difference in topological scores between the mutant sequence and the WT sequence, using the WT $C_\alpha$ structural coordinates. For a single point mutant, this difference in sequence-structure compatibility from WT is an overall measure of the environmental change resulting from the mutation at the residue position under consideration. At each residue position in the protein, a CMP score is defined as the mean of the residuals associated with all possible mutations (Fig. 3). Mathematically,

$$\begin{aligned}\mathrm{CMP}_j &= \frac{1}{20}\sum_{i=1}^{20}[(\text{mutant topological score})_{ij} \\ &\quad - (\text{wt topological score})] \\ &= \frac{1}{20}\sum_{i=1}^{20}(\text{mutant residual score})_{ij} \\ &= \{\text{mean residual score}\}_j \qquad (1)\end{aligned}$$

where the index $i$ refers to the 20 standard amino acids, and the index $j$ refers to the position in the primary sequence of the protein. The null residual associated with a substitution of a WT residue with itself is included at each position for completeness.

### Robustness With Respect To Amino Acid Point Representations of Proteins

The $C_\alpha$ coordinate point representation of amino acids in a protein for the purpose of Delaunay tessellation is clearly not unique. Use of the center of mass (CM) of amino acid side-chains as a means of discretizing the protein structure is one alternative that may provide a slight advantage by incorporating implicit information about side-chain directionality. Calculation of topological and residue environment scores, as well as 3D-1D and CMP profiles, for a protein represented by its amino acid CM coordinates requires that a scoring function first be generated based on Delaunay tessellation of a representative training set of proteins from the PDB using CM coordinates. The HIV-1 protease and T4 lysozyme results presented in this article reflect the use of both a CM coordinate representation of the protein structures as well as a CM coordinate-derived

```
           1        10        20        30          40           50        60               70        80         90
HIV-1      PQITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEM--SLPGRW---KPKMIGGIGGFIKVRQYDQIL---IEICGH----KAIGTVLVGPTPVN-IIGRNLLTQIGCTLNF
HIV-2      PQFSLWKRPVVTAYIEGQPVEVLLDTGADDSIVAGI--ELGNNY---SPKIVGGIGGFINTKEYKNVE---IEVLNK----KVRATIMTGDTPIN-IFGRNILTALGMSLNL
SIV        PQFSLWRRPVVTAHIEGQPVEVLLDTGADDSIVTGI--ELGPHY---TPKIVGGIGGFINTKEYKNVK---IEVLGK----RIKGTIMTGDTPIN-IFGRNLLTALGMSLNL
VISNA      PYVVTEAPPKIEARIGTTWKELLVDTGADRTIVRKHD-NTGIPK---GRIKLQGIGGIIEGEKWDQVE---IQYKGK---SIRGTLVVLPSSPVE-VLGRDNMGELGIGLIM
FIV        TTTTLEKRPEILIFVNGYPIKFLLDTGADITILNRRDFQVKNSIEN-GRQNMIGVGGGKRGTNYINVHLEIRDENYKTQCIFGNVCVLEDNSLIQPLLGRDNMIKFNIRLVM
EqiAV      VTYNLEKRPTTIVLINDTPLNVLLDTGADTSVLTTAHYNRLKYR---GRKYQGTGIGGVGGNVETFSTPVTIKKKGR----HIKTRMLVADIPVT-ILGRDILQDLGAKLVL
FELV       QGQDPPPEPRITLKIGGQPVTFLVDTGAQHSVLTRP--DGPLSD---RTALVQGATGSKNYRWTTDRR---VQLATG---KVTHSFLYVPECPYP-LLGRDLLTKLKAQIHF
Mo-MLV     QGQEPPPEPRITLKVGGQPVTFLVDTGAQHSVLTQN--PGSLSG---KSAWVQGATGGKRYRWTTDRK---VHLATG---KVTHSFLHVPDCPYP-LLGRDLLTKLKAQIHF
AKV-MLV    QGQEPPPEPRITLTVGGQPVTFLVDTGAQHSVLTQN--PGPLSD---RSAWVQGATGGKRYRWTTDRK---VHLATG---KVTHSFLHVPDCPYP-LLGRDLLTKLKAQIHF
M7-BEV     QGSGAPPEPRLTLSVGGHPTTFLVDTGAQHSVLTKA--NGPLSS---RTSWVQGATGRKMHKWTNRRT---VNLGQG---MVTHSFLVVPECPYP-LLGRDLLTKLGAQIHF
MMTV       VQEISDSRPMLHISLNGRRFLGLLDTGADKTCIAGRDWPANWPIHQ-TESSLQGLGMACGVARSSQPLRWQHEDKSG-----IIHPFVIPTLPFT-LWGRDIMKEIKVRLMT
HERV-K     ASQVSENRPVCKAIIQGKQFEGLVDTGADVSIIALNQWPKNWPKQK-AVTGLVGIGTASEVYQSTEILHCLGPDNQE-----STVQPMITSIPLN-LWGRDLLQQWGAEITM
RSV        VRVILTNTGSHPVKQRSVYITALLDSGADITIISEEDWPTDWPVMEAANPQIHGIGGGIPMRKSRDMIELGVINRDGSLERPLLLFPAVAMVRGS-ILGRDCLQGLGLRLTN
HTLV-1     ARRPVIKAQVDTQTSHPKTIEALLDTGADMTVLPIALFSSNTPL---KNTSVLGAGGQTQDHFKLTSLPVLIRLPFRTTPIVLTSCLVDTKNNWA-IIGRDALQQCQGVLYL
HTLV-2     QQQPILGVRISVMGQTPQPTQALLDTGADLTVIPQTLVPGPVKL---HDTLILGASGQTNTQFKLLQTPLHIFLPFRRSPVILSSCLLDTHNKWT-IIGRDALQQCQGLLYL
STLV       QKQPILHVQVSFSNTSPVNVQALLDTGADITVLPACLCPPDSNL---QDTTILGAGGPSANKFKILPRPVHIHLPFRRQPVTLTSCLIDINNQWT-ILGRDALQQCQGSLYL
BLV        RPSVAVYLSGPWLQPSQNQALMLVDTGAENTVLPQNWLVRDYPR---IPAAVLGAGGVSRNRYNWLQGPLTLALKPEGPFITIPKILVDTFDKWQ-ILGRDVLSRLQASISI
                                 *:*:**: : :                    :                                          : **: :     :
```

Fig. 4. Multiple sequence alignment of 17 retroviral proteases performed using ClustalW.[25] Numbering is based on the primary sequence of HIV-1 protease. Six residue positions marked by "*" are fully conserved; 10 residue positions marked by ":" are conserved when allowing for conservative substitutions (see text for details). All 16 conserved residue positions are highlighted in red in the ribbon diagram of the HIV-1 protease dimer of Figure 1(B). Abbreviations, as described in Pechik et al.,[24] are as follows: HIV, human immunodeficiency virus; SIV, simian immunodeficiency virus; VISNA, visna lentivirus; FIV, feline immunodeficiency virus; EqiAV, equine infectious anemia virus; FELV, feline leukemia virus; Mo-MLV, Moloney mouse leukemia virus; AKV-MLV, AKV mouse leukemia virus; M7-BEV, baboon endogenous virus M7; MMTV, mouse mammary tumor virus; HERV-K, human endogenous virus K; RSV, Rous sarcoma virus; HTLV, human T-cell leukemia virus; STLV, simian T-cell leukemia virus; BVL, bovine leukemia virus.

scoring function. The HIV-1 RT results, however, are derived using the original $C_\alpha$ approach. In the case of HIV-1 protease, the 3D-1D profiles obtained using the $C_\alpha$ and CM approaches are remarkably similar ($R^2 = 0.88$); the same is true for the CMP profiles of protease ($R^2 = 0.84$).

### Experimental Data

All of the activity levels of the mutants analyzed in the current work have been collected from previously published accounts detailing the results of large-scale experimental trials. In the first system of protein mutants, Loeb et al.[16] studied the ability of single missense mutations of HIV-1 protease to process the Pol precursor in an *Escherichia coli* expression system. The 336 mutants analyzed in the study were placed into three phenotypic categories. A mutant was considered to be positive if it maintained a level of activity similar to WT, as evidenced by the appearance of mature Pol products. The negative phenotype was reserved for mutants that did not yield any mature products, with only the unprocessed Pol being observed. Finally, the intermediate category contained a wide variety of phenotypes, with variable amounts of both mature and unprocessed Pol products being detected. The data set has subsequently been enlarged with the inclusion of 200 additional mutants (R. Swanstrom, personal communication), providing a total of 536 mutants for this study.

For the second system that we investigated, Rennell et al.[17] systematically introduced amber mutations into 163 of the 164 codons (all except for the initial AUG) of bacteriophage T4 lysozyme. The same 13 amino acid substitutions, representative of the entire spectrum of side-chain structural and chemical characteristics, were inserted at each of the 163 positions, yielding 2,015 single amino acid mutants of the enzyme as well as 104 nonmutants used as an experimental control. The effect of each mutation on T4 lysozyme function was qualitatively measured as high (++), medium (+), low (+/−), or negative (−) by visualizing the size of *E. coli* plaques that formed on agar plates. The authors provide a complete tabular listing of the activity levels of these mutant T4 lysozyme enzymes in their article. Given the substantial variability inherent in the experimental conditions, the analyses and results presented by the authors is based on the consideration of only two activity classes: positive/active (++ and + combined) and negative/inactive (+/− and − combined).

In the third experimental system, the phenotypic effects of 366 single amino acid substitutions, at residue positions comprising the fingers and palm subdomains of HIV-1 RT, were analyzed by Wrobel et al.[18] These subdomains cover a 109 amino acid segment of the 66-kDa subunit of the RT heterodimer that spans residues P95-E203. The level of RT mutant activity was quantitatively measured as a percentage of the activity displayed by a WT RT clone. Mutants were binned into four classes according to their activity levels: positive (≥50% of WT activity), high intermediate (20–50% of WT activity), low intermediate (3–20% of WT activity), and negative (<3% of WT activity).

## RESULTS AND DISCUSSION

### Evolutionarily Conserved Residue Positions Among Retroviral Proteases

We begin by performing a multiple sequence alignment on a set of 17 retroviral proteases, as described in Pechik et al.[24] Their study details a sequence alignment, partially based on a structural alignment, of a large set of aspartic proteases from a wide variety of organisms. Only the amino and carboxyl ends of the retroviral protease sequences are tabulated in their alignment, and the only information that we actually use from that study is the identity of the 17 retroviruses. Here, we begin de novo by locating the complete amino acid sequences of these retroviral proteases and focusing exclusively on a modified alignment of only these sequences, using ClustalW with an appropriately chosen set of parameters.[25]

As shown in Figure 4, the six fully conserved positions in the alignment, as they appear in the primary sequence for

**TABLE I. Residue Position Conservation and Sensitivity to Substitutions**

Contingency table based on number of residues:

| | S | NS | Total |
|---|---|---|---|
| C | 16 | 0 | 16 |
| NC | 48 | 35 | 83 |
| Total | 64 | 35 | 99 |

Contingency table based on number of substitutions:

| | N | I | P | Total |
|---|---|---|---|---|
| C | 111 | 6 | 3 | 120 |
| NC | 201 | 78 | 137 | 416 |
| Total | 312 | 84 | 140 | 536 |

A. Residues fully conserved in alignment (number of negative/intermediate/positive substitutions): Leu23 (3/1/0), Asp25 (3/0/0), Gly27 (7/0/0), Ala28 (4/0/0), Gly86 (16/0/0), Arg87 (11/0/0).

B. Residues conserved in alignment with the allowance for conservative substitutions: Leu24 (2/2/0), Thr26 (4/0/0), Asp29 (5/0/0), Thr31 (7/1/0), Leu33 (8/0/2), Gly49 (12/0/0), Ile84 (7/1/1), Asn88 (15/1/0), Leu90 (4/0/0), Leu97 (3/0/0).

C. Nonconserved residues: 83/99, with 48 of these leading to loss of enzyme activity through at least one substitution (201/78/137).

HIV-1 protease, are L23, D25, G27, A28, G86, and R87. An additional 10 positions contain only residues that are considered to be conservative substitutions of the protease residue. With residues grouped according to Dayhoff et al.[26] as (A, S, T, G, and P), (V, L, I, and M), (R, K, and H), (D, E, N, and Q), (F, Y, and W), and (C), a conservative substitution is defined as replacement with an amino acid from within the same class because they share similar structural or functional characteristics; an interclass substitution is considered to be nonconservative. The 10 positions in the alignment containing only conservative substitutions, as they appear in the primary sequence for HIV-1 protease, are L24, T26, D29, T31, L33, G49, I84, N88, L90, and L97. Collectively, these 16 positions (six fully conserved in the alignment, and 10 conserved with an allowance for conservative substitutions) will be referred to as the conserved residue positions of HIV-1 protease [Fig. 1(B)]. The remaining 83 positions in the alignment contain nonconservative substitutions of the HIV-1 protease residue and will be referred to as the nonconserved residue positions.

Although the experimental mutagenesis data obtained for HIV-1 protease and described in the Materials and Methods section is not exhaustive (536/1,881, or 28% of all possible single point mutants), a reasonable way to examine the functional significance of conserved residue positions in protease is to construct an association table measuring level of conservation of residue position versus sensitivity to mutations. As Table I illustrates, all 16 conserved positions are sensitive to at least one substitution (i.e., there exists at least one single point protease mutant at each of these positions with a negative phenotype), as are 48 of the 83 nonconserved positions. In both association tables, C and NC refer to the 16 conserved and 83 nonconserved residue positions. The S and NS abbreviations in the 2 × 2 association table refer to positions that display sensitivity to at least one substitution and positions that are not sensitive to any substitutions, respectively. Alternately, the N, I, and P designations in the 3 × 3 association table refer to the number of negative, intermediate, and positive mutants, respectively.

Applying a $\chi^2$ test statistic (based on the number of residues) with 1 degree of freedom (df) yields $\chi^2 = 10.44$ and results in rejection of the null hypothesis that no association exists between residue position conservation and level of sensitivity to mutation, with $P < 0.01$. The analysis was repeated by counting the total number of single point mutants at each level of activity (positive, intermediate, and negative phenotypes). Again, a $\chi^2$ test statistic (with 2 df) based on the number of substitutions yields $\chi^2 = 75.49$ and results in rejection of the null hypothesis, with $P < 0.001$.

### Mutagenesis at the Dimer Interface

An active area of research involves mutations at the HIV-1 protease dimer interface. A better understanding of these mutations may lead to the discovery of novel dimerization inhibitors for protease capable of inactivating the enzyme. Whereas the side-chains of Q2, T4, T96, and N98 are polar and directed outward toward the solvent, the residues P1, I3, L97, and F99 have side-chains directed inward toward the enzyme and form hydrophobic contacts.[27] For instance, F99 in one chain of the protease dimer is known to make extensive contacts with I3, V11, L24, I66, C67, I93, C95, and H96 in the complementary chain.[28] Using our computational geometry approach, these contacts are revealed by calculating the difference in 3D-1D profiles for a single subunit of an HIV-1 protease homodimer, with and without and F99A mutation in the complementary chain, as follows. First, the profile of the single chain is obtained following Delaunay tessellation of the complete WT protease dimer. Next, an F99A mutation is introduced in the opposite chain, and the profile of the chain under consideration is recalculated. The difference profile, shown in Figure 5, illustrates which residues form contacts with F99 in the complementary chain and to what degree the F99A mutation impacts their residue environment scores.

Using an *E. coli* expression system in which a tethered dimer of HIV-1 protease was coexpressed with β-galactosidase containing a protease cleavage site, Choudhury et al.[27] performed an alanine scan on the dimer interface residues. Residues were substituted with alanine individually and in pairs, in a single subunit as well as in both chains, and protease activity was measured by the frequency of β-galactosidase cleavage. Identical mutations in both subunits that diminished protease activity experienced modest to significant rebounds when one chain was

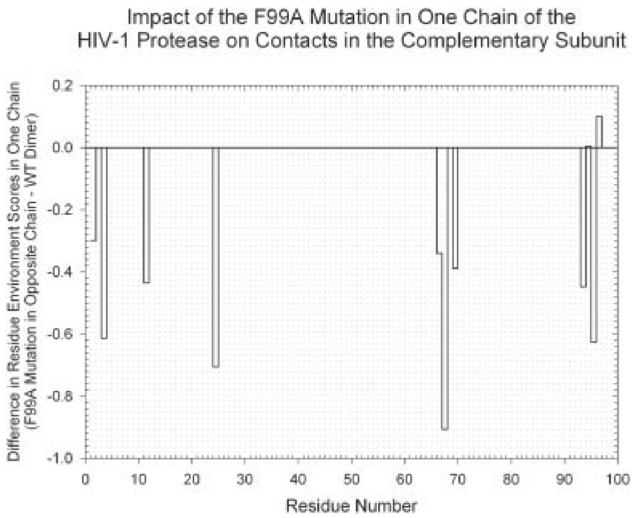


Fig. 5. Plot of the difference in 3D-1D profiles for one subunit in the HIV-1 protease homodimer with and without a single point mutation in the opposite chain. The WT profile is subtracted from the profile obtained when an F99A mutation is introduced in the complementary chain. Residues in the subunit exhibiting a non-zero difference participate in at least one nearest-neighbor Delaunay simplex with F99 of the opposite chain in the Delaunay tessellation of the protease homodimer.

maintained as WT. Plots of residual scores (the difference between the topological scores of the mutant and the WT proteases) versus % cleavage of β-galactosidase yields $R^2$ values of 0.61 and 0.57 for the homodimer and single-chain mutants of the tethered protease constructs, respectively, as illustrated in Figure 6 (A and B).

Comparing the residual scores for the double and single chain mutants in Figure 6(C) reveals that the individual substitutions showing the greatest increase by keeping one WT subunit are I3A, L97A, and F99A. These substitutions also lead to maximally negative residuals when introduced into either one or both subunits, echoing findings in previously published reports. Shultz and Chmielewski[28] obtained these results while analyzing an alanine scan of a prospective dimerization inhibitor of HIV-1 protease, and Bowman and Chmielewski[29] emphasize that L97 and F99 were found by Todd et al.[30] to be among the top three most important residues contributing Gibbs energy toward protease dimerization. However, alanine substitutions at the polar Q2 and T4 positions yield minimally negative residual scores at the N-terminus for protease when incorporated into either one or both subunits. The experiments of Choudhury et al.[27] reveal a minimal decrease in protease activity with these substitutions, and they refer to the minimal contribution of Q2 and T4 to the Gibbs energy of dimerization described in Todd et al.[30] Next, with the exception of the L97A+N98A substitution, the residual scores of the other paired mutations in both subunits of protease exhibit significant increases when a pair in one of the subunits is maintained as WT. This correlates well with the Choudhury et al.[27] data, whereby all four paired mutants exhibited increases in activity when one of the monomers did not incorporate the mutations. Finally, our results mirror previously reported findings concluding that, overall, the C-terminal residues tend to be more sensitive to substitutions than those at the N-terminus.[4,27,30,31]

## Structure-Function Correlation

Turning to the activity data for the 536 experimentally synthesized single point mutants of HIV-1 protease, we were interested in understanding the relationship between the level of activity of the mutants and their residual sequence-structure compatibility scores (difference between mutant and WT topological scores). At each level of activity (positive, intermediate, negative), a mean residual score was calculated by averaging the residuals of all the mutants in the given class. The mutants in each activity class were further subdivided into conservative (C) and nonconservative (NC) substitutions of the WT residue as described at the beginning of the Results section, and a mean residual score was also calculated for each subgroup.

The findings, summarized in Figure 7, reveal several important facts. First, a strong correlation exists between the level of mutant activity and the mean residual of the mutants in each class (ALL bar graph in Fig. 7). Next, an equally strong correlation exists between the level of activity and the mean residual for the NC substitutions in each class. This result appears to be intuitively clear, because NC mutations that do not adversely affect the functionality of HIV-1 protease are not expected to negatively impact the sequence-structure compatibility of the enzyme to the same extent as NC deleterious substitutions. Finally, no measurable correlation exists between the level of activity and the mean residual for the C substitutions in each class. Again, this follows intuitively from the fact that C substitutions, regardless of how they impact protease activity, are expected to minimally impact sequence-structure compatibility. Although C mutants exhibiting an intermediate level of activity appear to yield a more negative mean residual compared with their counterparts in the positive and negative classes, we consider this observation to be a consequence of both the qualitative nature of the activity measurements and the wide latitude for inclusion into the intermediate category.

These results parallel previous observations that were made as we deconstructed the strong inverse correlation found to exist between the 3D-1D and CMP profiles of HIV-1 protease ($R^2 = 0.88$).[15] As a way to begin to understand the factors influencing the correlation, the 3D-1D profile was compared with a CMP profile based on using only the C mutants, as well as a CMP profile obtained by using only the NC mutants. Whereas a C/NC clustering of the mutants for the purpose of obtaining their respective CMP profiles again led to a strong inverse correlation between the 3D-1D and NC-CMP profiles ($R^2 = 0.88$), no significant correlation existed between the 3D-1D and C-CMP profiles ($R^2 = 0.16$).[15] This earlier work on clustered CMP profiles reflecting the important role of the NC mutants is mirrored by the observations in Figure 7.

Next, residual scores were calculated for each of the 2,015 single point mutants of T4 lysozyme with experimen-

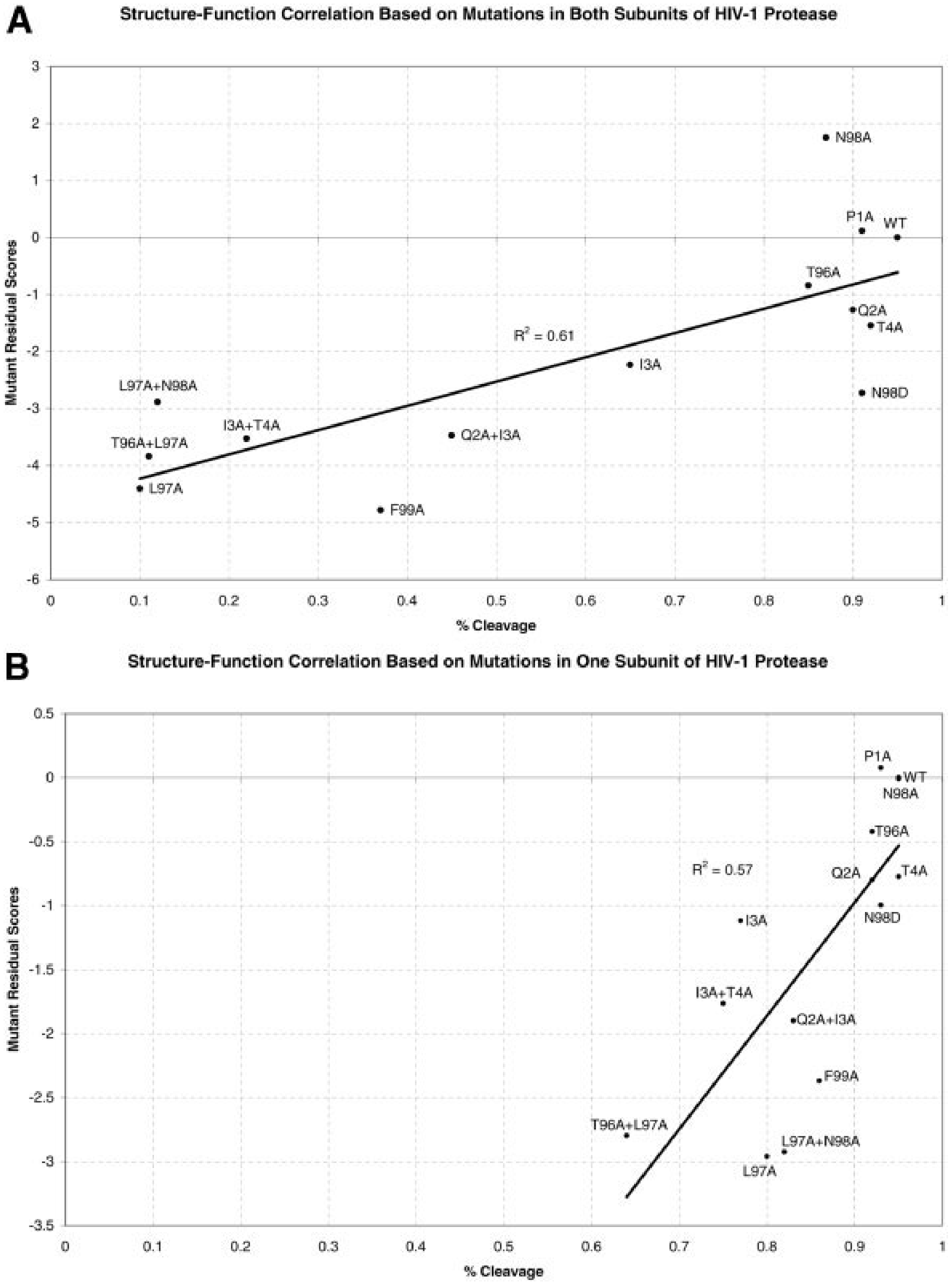


Fig. 6. Plots of residual score (the difference between the topological scores of the alanine mutant and the WT protease) versus % cleavage of β-galactosidase for the homodimer (**A**) and single-chain (**B**) mutants of tethered protease constructs. A strong structure-function correlation is evident in both cases, with $R^2 = 0.61$ in (A) and $R^2 = 0.57$ in (B). Residuals for single and double alanine mutants, occurring in a single chain as well as both subunits, are shown in (**C**).

tally measured activity levels, representing nearly 65% of the 164 residues × 19 substitutions/residue = 3,116 possible single point mutants. A structure-function correlation plot for T4 lysozyme analogous to that of HIV-1 protease in Figure 7 is shown in Figure 8(A), where the four activity classes used in the plot are identical to those defined by Rennell et al.[17] While the correlation is slightly tempered by the mean residual score of the negative class, we attribute this observation to the higher degree of activity measurement error for mutants in the medium and low classes compared with those in the high and negative classes, as described by the authors and leading them to use only two classes in their subsequent analyses of the data. However, Figure 8(A) again illustrates the overwhelming influence of the NC substitutions in driving the trend of the overall plot (ALL bar graph), with minimal mean residual scores for the C mutants that do not contribute to the trend. Following the approach of Rennell et al.,[17] Figure 8(B) presents the same data by considering only an active

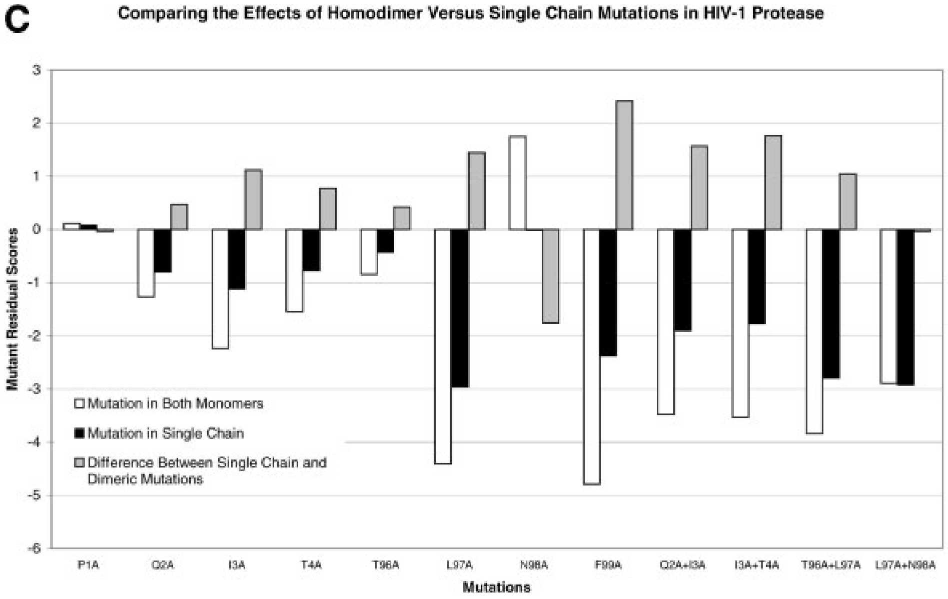
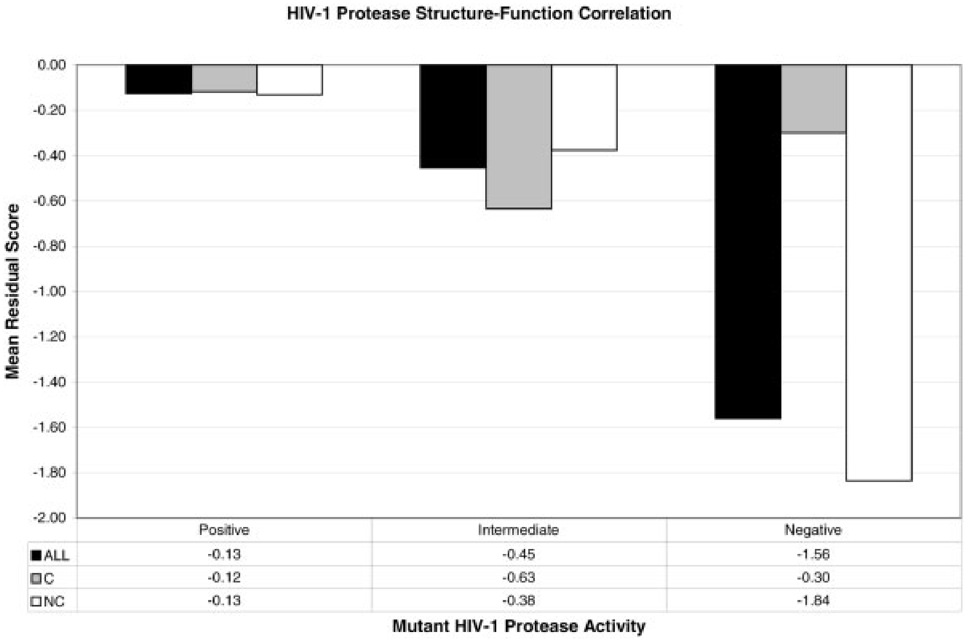


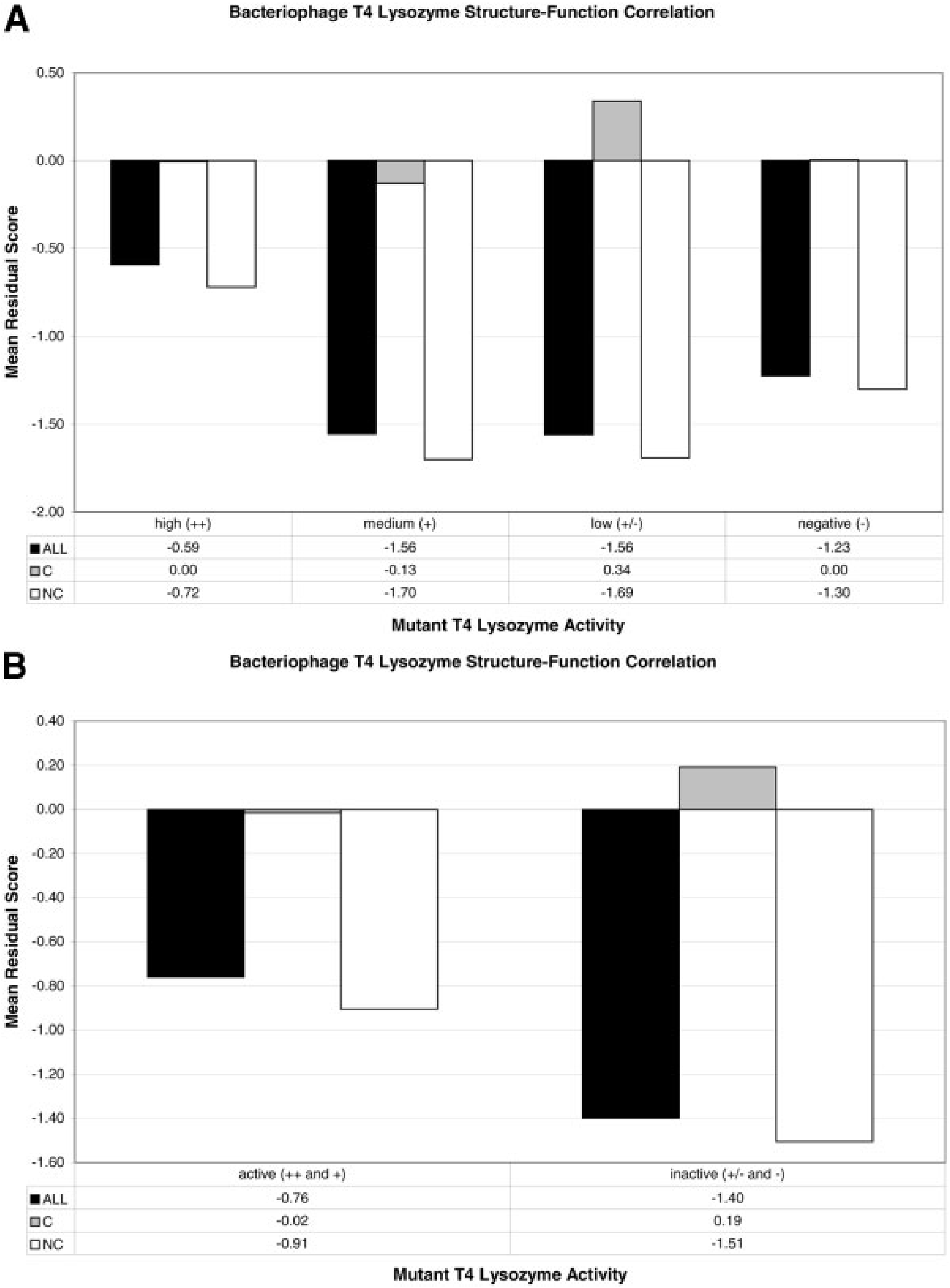
Figure 6. (Continued.)



| | Positive | Intermediate | Negative |
|---|---|---|---|
| ■ALL | -0.13 | -0.45 | -1.56 |
| ▨C | -0.12 | -0.63 | -0.30 |
| □NC | -0.13 | -0.38 | -1.84 |

Fig. 7. Comparison of the activity of 536 experimentally synthesized HIV-1 protease mutants with the mean residual of the mutants within each functional class. The residual of a mutant is defined as the difference between mutant and wild-type topological scores (i.e., sequence-structure compatibility scores). Two sample *t*-tests for independent samples with unequal variances show that there is a highly statistically significant difference between the mean residual scores of the positive and negative classes ($P = 1.65 \times 10^{-11}$), as well as the intermediate and negative classes ($P = 9.90 \times 10^{-6}$). Mutants in each activity class are further subdivided based on whether they are conservative (C) or nonconservative (NC) substitutions of the WT residue (see text for details). A significant structure-function correlation is evident, driven specifically by the NC mutants within the activity classes. The C mutants in each of the three activity classes generally have residuals that are small in magnitude because of the minimal change in sequence-structure compatibility from WT; hence, the mean of the residual scores of the C mutants in each activity class remains relatively constant across all three classes.

class (mutants in the high and medium classes combined) and an inactive class (mutants in the low and negative classes combined). With a two-class system, the correlation of mean residual score with activity is apparent, driven primarily by the NC substitutions with no contribution from the C mutants.



| | high (++) | medium (+) | low (+/-) | negative (-) |
|---|---|---|---|---|
| ■ALL | -0.59 | -1.56 | -1.56 | -1.23 |
| ▨C | 0.00 | -0.13 | 0.34 | 0.00 |
| □NC | -0.72 | -1.70 | -1.69 | -1.30 |

| | active (++ and +) | inactive (+/- and -) |
|---|---|---|
| ■ALL | -0.76 | -1.40 |
| ▨C | -0.02 | 0.19 |
| □NC | -0.91 | -1.51 |

Fig. 8. Comparison of the activity of 2,015 experimentally synthesized bacteriophage T4 lysozyme mutants with the mean residual of the mutants within each of four functional classes (**A**) and within each of two functional classes (**B**) as described in Rennell et al.[17] Two sample *t*-tests for independent samples with unequal variances show that there is a highly statistically significant difference between the mean residual scores of the high and medium classes ($P = 4.97 \times 10^{-9}$), the high and low classes ($P = 5.61 \times 10^{-5}$), and the high and negative classes ($P = 0.010$). Similarly, there is a highly statistically significant difference between the mean residual scores of the active and inactive classes ($P = 0.0003$). The C and NC designations are described in the caption to Figure 7 and in the text.

Finally, the calculated residual scores of the 366 single point mutants of RT with experimentally determined activity levels were used to produce the plots in Figure 9(A and B). These residue substitutions are restricted to 109 positions (P95-E203) comprising the fingers and palm subdomains of the 66-kDa subunit of the RT heterodimer, and they represent approximately 18% of all 109 residues × 19 substitutions/residue = 2,071 possible single point mutants within this region of the subunit. The plot in Figure 9(A) is based on the use of the four activity classes used by Wrobel et al.[18] in their work. As with the T4 lysozyme plot in Figure 8(A), the RT correlation of mean residual score with activity in Figure 9(A) is slightly tempered by the mean score of the negative class. How-

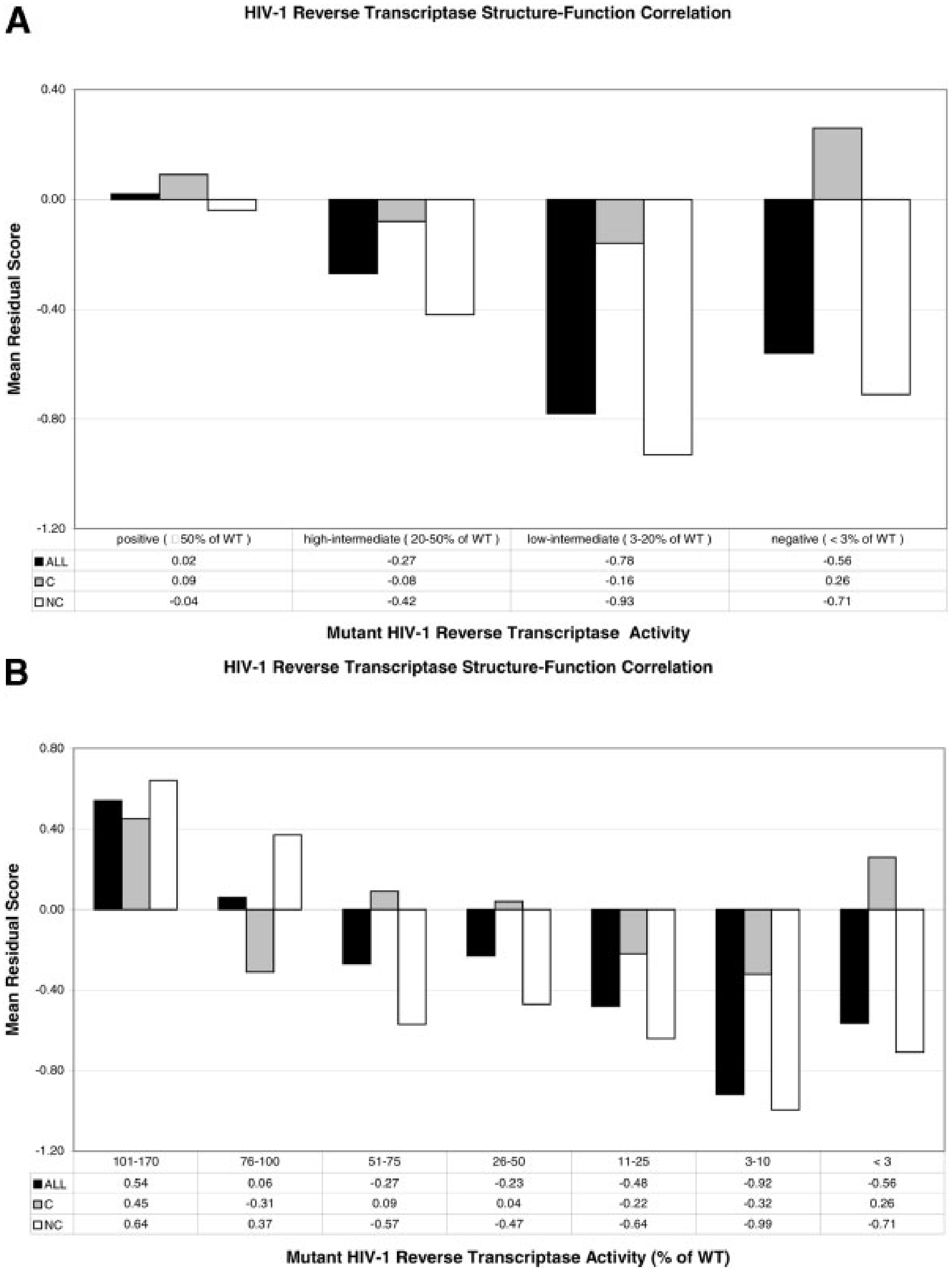


Fig. 9. Comparison of the activity of 366 experimentally synthesized HIV-1 RT mutants with the mean residual of the mutants within each of four functional classes (**A**), as described in Wrobel et al.,[18] and within each of seven functional classes (**B**). Two sample *t*-tests for independent samples with unequal variances show that there is a highly statistically significant difference between the mean residual scores of the positive and negative classes ($P = 0.013$). The C and NC designations are described in the caption to Figure 7 and in the text.

ever, this observation is likely attributable to the fact that so few mutants have been experimentally analyzed within the fingers and palm subdomains for inclusion into the plot. Figure 9(B) shows that this correlation is essentially maintained after increasing the number of activity classes from four to seven. The plots in Figure 9(A and B) also reveal that the correlation of mean residual score with activity is again driven primarily by the NC mutants, with no effective contribution by the mean residual scores of the C mutants in each class.

The overwhelming similarity of the results for the three model systems described above, especially given the limitations on number of single point mutants with known activity in each system available for the analysis, supports the suggestion that mutant residual scores encapsulate structural information about proteins that can be used to illuminate the strong impact of protein structure on function. Additionally, whereas the residual scores of the NC protein mutants are specifically correlated with changes in activity, the residual scores of the C mutants are generally small in magnitude because of the minimal impact that C substitutions typically have on sequence-structure compatibility.

## CONCLUSIONS

A multiple sequence alignment of retroviral aspartic proteases has identified six fully conserved positions in the 99 residues comprising HIV-1 protease, and an additional 10 positions are conserved when allowing for conservative substitutions. Comparing this information with the results of a study that measured the activity of 536 experimentally synthesized single point HIV-1 protease mutants reveals a strong association between residue position conservation and sensitivity to mutation. Data from a second study examining the activity of protease mutants obtained via an alanine scan of the dimer interface residues is compared with topological score data for the mutants (an indicator of sequence-structure compatibility). Specifically, a significant correlation is observed between the activity levels and the residuals (the change in sequence-structure compatibility from WT) of the mutants. The residuals associated with these mutants are less dramatic when the mutants occur in a single chain as opposed to both subunits of the HIV-1 protease homodimer. Additionally, our results echo previous findings that L97 and F99 are two of the three greatest contributors of Gibbs energy for HIV-1 protease dimerization, Q2 and T4 contribute the least toward the Gibbs energy, and C-terminal residues are more sensitive to mutations than those at the N-terminus.

A comparison is also made between the measured activity of the 536 experimentally synthesized HIV-1 protease mutants of the first study and their corresponding residual scores, again yielding a strong structure-function correlation. Specifically, each mutant is placed in one of three functional classes (positive, intermediate, or negative) based on its level of activity relative to WT. At each activity level, a mean residual is calculated for the set of mutants in that class. We observe that a decrease in activity level is associated with a decrease in mean residual. This correlation is driven primarily by the nonconservative residue mutations in each of the activity classes. Within each activity class, the mean residual of the conservative substitutions is minimal in magnitude, and they do not contribute to the correlation. Analogous results are also observed with 2,015 mutants of bacteriophage T4 lysozyme as well as 366 mutants of HIV-1 RT. These results support our assertion that protein topological scores based on the Delaunay tessellation encode the necessary structural information required to demonstrate the structure-function relationship inherent in proteins.

## ACKNOWLEDGMENTS

This work was supported in part by the PhRMA Faculty Development Award in Bioinformatics (1998–2000). The authors thank Ron Swanstrom, Clyde Hutchison, and John Wrobel for the experimental data on HIV-1 protease and HIV-1 RT mutant activities and for helpful discussions. The authors also thank Lakshmi Kumar Matukumalli for preliminary data collection on T4 lysozyme mutants, Todd Taylor for help with the four-body potentials, and Andrew Carr for help with creating the graphical image of tessellated HIV-1 protease. Finally, the ribbon diagram of the HIV-1 protease dimer was produced using the UCSF Chimera package from the Computer Graphics Laboratory, University of California, San Francisco (supported by NIH P41 RR-01081).

## REFERENCES

1. Wlodawer A, Vondrasek J. Inhibitors of HIV-1 protease: a major success of structure-assisted drug design. Annu Rev Biophys Biomol Struct 1998;27:249–284.
2. Shao W, Everitt L, Manchester M, Loeb DD, Hutchison CA 3rd, Swanstrom R. Sequence requirements of the HIV-1 protease flap region determined by saturation mutagenesis and kinetic analysis of flap mutants. Proc Natl Acad Sci USA 1997;94(6):2243–2248.
3. Strisovsky K, Tessmer U, Langner J, Konvalinka J, Krausslich HG. Systematic mutational analysis of the active-site threonine of HIV-1 proteinase: rethinking the "fireman's grip" hypothesis. Protein Sci 2000;9(9):1631–1641.
4. Ishima R, Ghirlando R, Tozser J, Gronenborn AM, Torchia DA, Louis JM. Folded monomer of HIV-1 protease. J Biol Chem 2001;276(52):49110–49116.
5. Kumar M, Hosur MV. Adaptability and flexibility of HIV-1 protease. Eur J Biochem 2003;270(6):1231–1239.
6. Zoete V, Michielin O, Karplus M. Relation between sequence and structure of HIV-1 protease inhibitor complexes: a model system for the analysis of protein flexibility. J Mol Biol 2002;315(1):21–52.
7. Mahalingam B, Louis JM, Hung J, Harrison RW, Weber IT. Structural implications of drug-resistant mutants of HIV-1 protease: high-resolution crystal structures of the mutant protease/substrate analogue complexes. Proteins 2001;43(4):455–464.
8. Mahalingam B, Boross P, Wang YF, et al. Combining mutations in HIV-1 protease to understand mechanisms of resistance. Proteins 2002;48(1):107–116.
9. Chasman D, Adams RM. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. J Mol Biol 2001;307(2): 683–706.
10. Karchin R, Diekhans M, Kelly L, et al. LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. Bioinformatics 2005;21(12):2814–2820.
11. Krishnan VG, Westhead DR. A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function. Bioinformatics 2003;19(17):2199–2209.
12. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res 2003;31(13):3812–3814.
13. Verzilli CJ, Whittaker JC, Stallard N, Chasman D. A hierarchical Bayesian model for predicting the functional consequences of amino acid polymorphisms. Appl Statistics 2005;54:191–206.
14. Wang Z, Moult J. SNPs, protein structure, and disease. Hum Mutat 2001;17(4):263–270.
15. Masso M, Vaisman II. Comprehensive mutagenesis of HIV-1 protease: a computational geometry approach. Biochem Biophys Res Commun 2003;305(2):322–326.
16. Loeb DD, Swanstrom R, Everitt L, Manchester M, Stamper SE, Hutchison CA 3rd. Complete mutagenesis of the HIV-1 protease. Nature 1989;340(6232):397–400.
17. Rennell D, Bouvier SE, Hardy LW, Poteete AR. Systematic mutation of bacteriophage T4 lysozyme. J Mol Biol 1991;222(1):67–88.
18. Wrobel JA, Chao SF, Conrad MJ, et al. A genetic approach for identifying critical residues in the fingers and palm subdomains of HIV-1 reverse transcriptase. Proc Natl Acad Sci USA 1998;95(2): 638–645.
19. Singh RK, Tropsha A, Vaisman II. Delaunay tessellation of

proteins: four body nearest-neighbor propensities of amino acid residues. J Comput Biol 1996;3(2):213–221.

20. Vaisman II, Tropsha A, Zheng W. Compositional preferences in quadruplets of nearest neighbor residues in protein structures: statistical geometry analysis. Proceedings of the IEEE Symposia on Intelligence and Systems 1998:163–168.
21. Zheng W, Cho SJ, Vaisman II, Tropsha A. A new approach to protein fold recognition based on Delaunay tessellation of protein structure. Pac Symp Biocomput 1997:486–497.
22. Barber CB, Dobkin DP, Huhdanpaa HT. The quickhull algorithm for convex hulls. ACM Trans Math Software 1996;22:469–483.
23. Berman HM, Westbrook J, Feng Z, et al. The Protein Data Bank. Nucleic Acids Res 2000;28(1):235–242.
24. Pechik IV, Kashparov IV, Novikova LA, Andreeva NS. Comparison of amino acid sequences and three-dimensional structures of aspartic proteases for use in their protein engineering. Mol Biol 1999;33(3):491–502.
25. Chenna R, Sugawara H, Koike T, et al. Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res 2003;31(13):3497–3500.
26. Dayhoff MO, Schwartz RM, Orcut BC. A model for evolutionary change in proteins. In: Dayhoff MO, editor. Atlas of protein sequence and structure. Vol 5. Washington, DC: National Biomedical Research Foundation; 1978. p 345–352.
27. Choudhury S, Everitt L, Pettit SC, Kaplan AH. Mutagenesis of the dimer interface residues of tethered and untethered HIV-1 protease result in differential activity and suggest multiple mechanisms of compensation. Virology 2003;307(2):204–212.
28. Shultz MD, Chmielewski J. Probing the role of interfacial residues in a dimerization inhibitor of HIV-1 protease. Bioorg Med Chem Lett 1999;9(16):2431–2436.
29. Bowman MJ, Chmielewski J. Novel strategies for targeting the dimerization interface of HIV protease with cross-linked interfacial peptides. Biopolymers 2002;66(2):126–133.
30. Todd MJ, Semo N, Freire E. The structural stability of the HIV-1 protease. J Mol Biol 1998;283(2):475–488.
31. Louis JM, Ishima R, Nesheiwat I, et al. Revisiting monomeric HIV-1 protease. Characterization and redesign for improved properties. J Biol Chem 2003;278(8):6085–6092.
32. Pettersen EF, Goddard TD, Huang CC, et al. UCSF Chimera: a visualization system for exploratory research and analysis. J Comput Chem 2004;25(13):1605–1612.