



LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

StralSV: assessment of sequence variability within similar 3D structures and application to polio RNA-dependent RNA polymerase

A. Zemla, D. Lang, T. Kostova, R. Andino, C. Zhou

December 6, 2010

BioMed Central Ltd

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

For Submission to: BMC Bioinformatics

StralSV: assessment of sequence variability within similar 3D structures and application to polio RNA-dependent RNA polymerase

Adam T Zemla*, Dorothy M Lang, Tanya Kostova, Raul Andino, and Carol L Ecale
Zhou*

Global Security Computing Applications Division, Lawrence Livermore National
Laboratory, Livermore, CA USA 94550

National Science Foundation, Arlington, VA USA 22230

University of California, San Francisco, CA USA 94143

*Corresponding authors

Email addresses:

ATZ: zemla1@llnl.gov

DML: lang21@llnl.gov

TK: tvassile@nsf.gov

RA: raul.andino@ucsf.edu

CEZ: zhou4@llnl.gov

Abstract

Background. Most of the currently used methods for protein function prediction rely on sequence-based comparisons between a query protein and those for which a functional annotation is provided. A serious limitation of sequence similarity-based approaches for identifying residue conservation among proteins is the low confidence in assigning residue-residue correspondences among proteins when the level of sequence identity between the compared proteins is poor. Multiple sequence alignment methods are more satisfactory—still, they cannot provide reliable results at low levels of sequence identity. Our goal in the current work was to develop an algorithm that could overcome these difficulties and facilitate the identification of structurally (and possibly functionally) relevant residue-residue correspondences between compared protein structures.

Results. Here we present StralSV, a new algorithm for detecting closely related structure fragments and quantifying residue frequency from tight local structure alignments. We apply StralSV in a study of the RNA-dependent RNA polymerase of poliovirus and demonstrate that the algorithm can be used to determine regions of the protein that are relatively unique or that shared structural similarity with structures that are distantly related. By quantifying residue frequencies among many residue-residue pairs extracted from local alignments, one can infer potential structural or functional importance of specific residues that are determined to be highly conserved or that deviate from a consensus. We further demonstrate that considerable detailed structural and phylogenetic information can be derived from StralSV analyses.

Conclusions. StralSV is a new structure-based algorithm for identifying and aligning structure fragments that have similarity to a reference protein. StralSV analysis can be

used to quantify residue-residue correspondences and identify residues that may be of particular structural or functional importance, as well as unusual or unexpected residues at a given sequence position.

Background

The ability to calculate accurate sequence alignments between related proteins is critical for many bioinformatics applications that involve comparative analysis. Derived from calculated alignments, residue-residue correspondences allow construction of sequence motifs and profiles important in building homology models or in predicting protein functions. Most of the currently used methods for protein function prediction rely on sequence-based comparisons between a query protein and those for which a functional annotation is provided. A serious limitation of sequence similarity-based approaches for identifying residue conservation among proteins is the lack of, or very low, confidence in assigning residue-residue correspondences among proteins when the level of sequence identity between the compared proteins is poor. Indeed, it was shown by Rost (1999) that more than 95% of all pair-wise alignments occurring in the so-called twilight zone (20-35% sequence identity) may be incorrect (Chodanowski 2008). Multiple sequence alignment methods are more satisfactory—still, they cannot provide reliable results at low levels of sequence identity, especially if the number of available closely related proteins is small (i.e., when the protein family has rather few members, or the list of related proteins that has been identified is short).

Having 3D structural information for a given protein can be especially useful in deriving functional annotation (Yakunin 2004). Structure comparison algorithms provide

much higher confidence in assignment of residue-residue correspondences than do sequence-based algorithms. Nevertheless even calculated structural alignments may be inaccurate: for some compared proteins, or regions therein, more than one possible superposition can reasonably be reported, and it may be difficult to decide which alignment is most satisfactory (Godzik 1996, Mayr 2007). Rigid body comparisons on the chain level have limitations when comparing multi-domain proteins with different conformations between domains. Comparisons on the domain level may yield better results, but splitting of structures into domains can be problematic, and there is no reliable method that can do this automatically. Even within compared structural domains, significant deviations can be observed in some loop regions, or due to large insertions or different conformations of structural motifs, all of which can significantly affect detection of structural residue-residue correspondences when rigid body approaches are used for alignment calculations. Several algorithms have been proposed to facilitate flexible protein structure alignment calculations (Shatsky 2002, Ye 2003, Menke 2008), but the complexity of such calculations remains a challenging development goal for many researchers. Another difficulty in identifying similar regions in compared protein structures lies in the possibility that analogous regions in structurally related proteins may display differences in sequential ordering of the motifs due to circular permutations or convergent evolution (Lindqvist 1997, Milik 2003). The majority of the existing flexible protein structure alignment algorithms report only sequential alignments, and there are very few (with varying levels of success) that can detect and align structures between which there are differences in the ordering of their structure motifs (Kolbeck 2006, Dundas 2007, Salem 2010).

The accuracy of calculated structural alignments can also depend on the nature of compared structural models. The atomic coordinates obtained from experimentally solved structures (x-ray crystallography or nuclear magnetic resonance spectroscopy) are always associated with some degree of uncertainty resulting from experimental errors from the intrinsic flexibility of the proteins or from atom vibrations (Domingues 2004). Such structural deviations may sometimes significantly affect the calculated alignments and lead to incorrect conclusions about sequence motifs, profiles, or possible residue substitutions within analyzed functional regions in proteins. The accuracy of calculated residue-residue correspondences can be improved by refinement methods that evaluate results produced by different structure-based alignment programs or explore sequence-based alignments using, for example, the Conserved Domain Database (CDD) as a set of reference alignments (Kim 2009).

Our goal in the current work was to develop an algorithm that could overcome these difficulties and facilitate the identification of structurally (and possibly functionally) relevant residue-residue correspondences between compared protein structures. Our approach is to first detect similar structural motifs, and consequently derive structure-based alignments from the calculated local superpositions of corresponding similar regions. Our StralSV algorithm detects structurally similar regions within a given pair of protein structures, and reports residue-residue correspondences only from those local regions that are contained within a larger, similar structural context. When for a given reference structure a structure-based search is performed on a set of proteins from the Protein Data Bank (PDB), StralSV identifies all structurally similar fragments from that set, evaluates the calculated structure-based alignments between the query (reference)

motif (designated “segment” in this work) and the detected structure fragments, and quantifies the observed sequence variability at each residue position on the query structure. Here we describe how the StralSV algorithm works, and we apply StralSV in a study of the RNA-dependent RNA polymerase of poliovirus.

Methods

Description of the StralSV algorithm.

StralSV is an algorithm that identifies protein structure fragments having a 3D structure similar to that of a query structure, performs structure-based alignments between the query and the fragments, and quantifies at each position along the query structure the sequence variability represented among the selected fragments relative to the query. StralSV takes as input a query structure of interest, a database of protein structures, and various parameters (discussed below) that control the selection of fragments from the database and the sequence variability calculations. Fig. 1 illustrates the steps in the algorithm. The algorithm uses a sliding-window approach for breaking the query structure into overlapping segments, each of which is independently used to identify from the input database protein structure fragments with 3D similarity. A recommended (default) `window_size` parameter is set to 90 amino acids in length, although an arbitrary length can be chosen. The query structure is thus split into overlapping segments of length `window_size`; overlaps are 1/2 the length of the `window_size`. A final segment is taken one `window_size` in length extending from the C-terminus to ensure that all portions of the query structure are represented within a segment no shorter than the `window_size`. Each so-calculated query segment is then compared to all protein

structures in the database using the LGA (local-global alignment) code (Zemla 2003) to identify structure fragments with sufficient structure similarity to the query segment. LGA uses a cutoff parameter (LGA_S), which is a measure of structure similarity based on an optimal alignment. The value of LGA_S can be set by the user, although a value of at least 55% ensures that there is good structural similarity between the segment and the fragment over more than half the length of the segment. LGA uses the GDT (global distance test) and LCS (longest continuous segment) algorithms to calculate optimal alignments between structures. LGA also requires a distance cutoff parameter, which determines the maximum distance allowed between alpha carbons for a given alignment; typically this parameter is set to 4.0 Å corresponding to window_size values of 90 residues. Thus, fragments with structure similarity to the query segment of at least LGA_S = 55% are identified. Each fragment is then evaluated to determine the tightness of its alignment to the query segment.

The criteria for tight structure similarities in local regions (spans), described in Zemla et al. (2007), are used to identify ranges within the alignment that have tight local superpositions. Each residue-residue pair (closest aligned single residues from the query segment and database fragment) is assigned a score by calculating the local RMSD (root mean square deviation) among the surrounding residue-residue pairs. A continuous set of residue-residue pairs that fulfil the RMSD cutoff of 0.5 Å comprises a span. A minimum size of the calculated span (span_size) is used as an input parameter to StralSV, and is typically specified as 3, 5, or 7, but can be of arbitrary length. For the work reported here, we selected span_size = 5 and minimum span length (shortest acceptable local alignment without gaps), based on previous experience (Zemla et al.

2007). From each alignment is extracted a set of spans. All alignments that contain at least one span of length no less than the specified minimum span length are deemed “qualified hits”. (For an illustration of a “span”, see Suppl. Fig. 1.) All residue-residue pairs that are contained within a span are used to calculate the sequence variability data at the corresponding position in the query structure. Note that not all residues from calculated structure alignments contribute to the variability statistics at a given position in the query segment; regions in which the local RMSD distances between corresponding residues exceed 0.5 Å induce breakpoints between spans. Also, because the algorithm uses overlapping segments, duplications are appropriately factored out in calculating the sequence variability.

A frequency matrix (Table 1A) is constructed for each position in the query structure by tallying the frequency at which each amino-acid (the 20 standard amino-acids plus ‘X’, corresponding to unusual or modified residues) is observed within the spans. From this matrix are extracted statistics describing the number of positional hits (residue-residue correspondences contributing sequence variability data) per position and a list of residues observed at each position, ordered by frequency of occurrence. These statistics are used to construct a variability profile (Table 1B), which can be used to identify positions at which there is relatively high or low sequence variability in structure context. The sample profile given in Table 1B shows the observed residues for positions 229 through 269 for polio RNA-dependent RNA polymerase (RdRp) run as the query protein against the complete PDB database (released on November 24, 2009).

Selection of parameters for StralSV using benchmark structure set.

To illustrate the StralSV algorithm, we selected poliovirus RdRp because poliovirus has been extensively studied and its polymerase is a member of a widely distributed protein family. We selected an RMSD span_size of 5 and minimum span length of 5 for the analysis presented here based on previous work (Zemla et al. 2007), which suggests that 5 is a reasonable minimum length over which to impose a tight alignment; 3 is the minimum value for span_size that is meaningful (since any 2 residues can be perfectly aligned), and 7 imposes stringency that tends to eliminate capture of some related (desired) structure fragments. In this work, part of our effort involved determining suitable parameters for window_size and distance_cutoff in analyzing poliovirus RdRp. As window_size is increased, the stringency with which similar structure fragments are selected is increased, due to the greater structural context that is provided by the query segment. As distance cutoff is increased, that stringency is reduced, because more laxity in the alignment is allowed. Thus, for larger window_size values, a larger distance cutoff should be applied in order for the algorithm to not eliminate related structure fragments that have local structural deviations with respect to the query segment. Likewise, as the window size is decreased, the distance cutoff should be reduced as well, to prevent capture of many small, unrelated structure fragments. To determine what combinations of window_size and distance_cutoff values would provide comparable results, we performed a test (Fig. 2) involving the capture of structure fragments from a benchmark data set containing 38 polymerase structures plus more than 1000 other structures randomly selected from the PDB. Suitable parameters input to StralSV were expected to capture only structurally related fragments (i.e., the polymerases), whereas overly stringent parameters were expected to yield result sets

lacking at least some of the polymerase fragments, and low stringency parameters were expected to capture unrelated fragments. In this way, we examined the dependency between `window_size` and `distance_cutoff` values to determine optimal parameter settings for StralSV.

We conducted a test whereby `window_size` and `distance_cutoff` parameters were varied from 20 to 150 and 1.0Å to 6.5Å, respectively (Fig. 2). We ran StralSV using each `window_size/distance_cutoff` combination for all query (poliovirus RdRp) segments enclosing residue G64. We selected a region in the N-terminal portion of the protein upon which to focus this exercise. This region defined by residues 1-140 tends to be structurally unique and contains a variety of secondary structure elements. Furthermore, we selected those segments containing residue G64 only because of its significance as a functional residue (Fig. 2A; Gohara 2000). We determined how many qualified hits were obtained for each `window_size/distance_cutoff` combination (Fig. 2B). Parameter settings comprising `window_size` values ranging from 70 to 90 at distance cutoffs 3.0Å to 5.0Å gave satisfactory results; fragments from the seeded 38 polymerase structures were reliably captured within a set of `window_size` and `distance_cutoff` parameter value combinations within these ranges (dotted oval in Fig. 2B). `Window_size` values smaller than 70 tended to yield qualified hit sets missing some of the 38 polymerases (i.e., true positives) for lesser (more stringent) values of `distance_cutoff` and tended to capture increasing numbers of unrelated (false positive) structure fragments as the distance cutoff was increased. Also, `window_size` values smaller than 70 were sensitive to the distance cutoff value, yielding acceptable result sets only for narrow ranges of distance cutoff. Very large `window_size` values (e.g., 150) resulted in selection of all true

positives for all but the very tightest (distance_cutoff <2.5Å) alignments. The parameter settings highlighted in red type in Fig. 2B were selected as default values for StralSV, which is provided as a web service at <http://as2ts.llnl.gov/stralsv/>.

Analysis of poliovirus RdRp.

Parameters. Based on parameters applied in Zemla et al. 2007 and the analysis described above, we used StralSV to analyze sequence variability in structure context for poliovirus RdRp (PDB: 1rdr; Hansen 1997) using RMSD span_size = 5, minimum span length of 5, LGA_S cutoff 55%, and window_size/distance_cutoff combinations 50/2.5Å, 70/3.5Å, 80/4.0Å, and 90/4.0Å.

Plotting of StralSV results. StralSV produced variability matrixes and sequence profiles (not shown; for examples see Tables 1 A, B) from which were extracted data for analyzing related structure fragments and quantifying positional variability. Plots containing data extracted directly from the matrix and profiles files are shown in Figs. 3-

5. We performed additional analyses in order to annotate the primary StralSV results:

1) Secondary structure predictions were calculated for poliovirus RdRp using DSSP (Rost and Sander 1993) and were plotted along the x-axis of Fig. 3. DSSP output was simplified as follows: helix, comprising alpha helix, pi helix, and 3/10 helix; strand, comprising extended strand or residue in isolated beta bridge; and coil, comprising hydrogen bonded turn, bend, or any other. 2) To determine whether we could discern patterns with respect to positional hit frequency and previously identified polymerase

sequence motifs, we overlaid a plot comprising qualified hits versus sequence position with the positions of the well known motifs, extracted from the literature (Fig. 5, gray boxes with labels A-G). This was accomplished by examining sequence alignments and extracting coordinates defining the known sequence motifs A through G from papers that included poliovirus RdRp among the aligned sequences (Hansen 1997, Bruenn 2003, O'Reilly 1998, Xu 2003); Because there was considerable inconsistency regarding the boundaries of the sequence motifs, we defined inclusive boundaries for each motif whereby residues were included if they were identified within a motif in any of the sequence alignments reported in the literature. 3) In order to inspect the most abundant sequence variants (i.e., the most conserved positions) in the context of the positional hit frequency along the reference protein, we extracted from the matrix files the frequencies of the most frequent residues and plotted them versus sequence position along with the positional hits for window size 80 (Fig. 5). 4) The literature was searched for evidence of functional annotation for the most frequently observed residue positions (Suppl. Table 1) and positions for which functional annotation was identified were marked in Fig. 5.

Annotation of representative positional hits. The qualified hits for six selected sequence positions (positional hits) of polio RdRp (positions identified in Fig. 3) were categorized and quantified based on SCOP (version 1.75, June 2009 release) identifiers (Table 2 and Suppl. Tables 2-7). Note that because SCOP is a manually curated database of structure domains from the PDB, there is some delay (about 18 months) before a new PDB entry is classified in SCOP. Therefore, we have included in Table 2 and Suppl. Tables 2-7 data pertaining to qualified hits and sequence variability based

on StralSV analysis using PDB in addition to those hits and variabilities that could be categorized by SCOP classification. The hits were evaluated separately for each of windows 50, 70, 80 and 90, at six sequence positions selected to be representative of the types of frequency variation (of qualified hits) over the length of the polymerase (see specified positions in Fig. 3). For each of the six selected positions, the following information was extracted from the StralSV matrix output file: all PDB templates (including chain information) that contributed to the profile at that position, the corresponding LGA_S score, the sequence identity and number of (template) amino acids that matched the query, and the amino acid of the template. The SCOP identifier for each template was determined from the PDB annotation.

Results

Effects of parameter settings on capture of database structure fragments.

Numbers of positional hits (residues within spans) corresponding to each position along the poliovirus query structure were plotted for window_size/distance_cutoff parameter pairs 70/3.5A, 80/4.0A, and 90/4.0A (hereafter referring only to the window_size) (Fig. 3). In some regions, there was observed considerable variation in the numbers of positional hits captured for the different window_size settings (e.g., positions 290-312 and 390-440), whereas at other positions, window_size had little effect on the numbers of qualified hits captured (most notably positions 365-380). However, with the exception of positions 425 and 145-150, when the number of positional hits reached a minimum, the identification of positional hits was robust and did not depend on the window size. Observations of positional hit frequency variations over the length of poliovirus RdRp

led to identification of a variability profile (illustrated in Figs. 3 and 5). The regions of high variability were “locked” between regions of high specificity, which exhibited little or no change with the variation of window_size. As window_size was increased, the numbers of positional hits tended to decrease in those regions for which window_size-dependent variability was observed. Comparison of positional hit variabilities to secondary structure (Fig. 3, along x-axis) revealed that the window-related variability tended to occur in regions of helical secondary structure. A polymerase “baseline”, amounting to fewer than 50 positional hits (corresponding primarily to polymerase structures) was observed along the entirety of the poliovirus chain (Fig. 3 x-axis). All polymerase baseline regions, with the exception of region 410-463, were identifiable independently of window_size. The latter C-terminal region showed high specificity at window size 90. In this region, lowering the window_size value to just 70 resulted in the capture of many (>900) structure fragments that have apparent structural similarity to poliovirus RdRp (Fig. 3, tall red peaks at right side of plot). The N-terminal region of the poliovirus protein yielded the least positional hits, indicating relative structural uniqueness in this region.

Composition of captured structure fragments at specific positions. We selected six positions (Fig. 3, residues numbered at top of plot) at which to examine the diversity of structure fragments captured at each of four window_size values (Table 2, Suppl. Tables 2-7). Orientations within the structure of poliovirus RdRp of these six positions are indicated on the structure model shown in the inset of Fig. 3. Qualified hits were examined for positions at which numbers of positional hits were relatively invariant (D238, D328, L374) or highly variant (N297, H398, H413) among the calculations run at

the four `window_size` values. (In some cases preferences for selection of a particular residue for this analysis was based on existence of a functional annotation for that residue and, therefore, of biological interest (e.g., D238; Hansen 1997)). We included `window_size` 50 in this analysis in order to determine what effect a very small (low stringency) `window_size` value might have on the structure-function diversity of qualified hits (i.e., would it greatly increase the diversity?). For each position and each `window_size`, qualified hits were categorized by SCOP concise classification strings (Murzin et al. 1995). Hits were categorized into four families in the e.8.1 (DNA/RNA polymerases) superfamily: e.8.1.1 (DNA polymerase I), e.8.1.2 (Reverse transcriptase), e.8.1.4 (RNA-dependent RNA-polymerase), and e.8.1.6 (dsRNA phage RNA-dependent RNA-polymerase), and into various other non-e.8 SCOP classifications.

At positions D238, D328, and L374, qualification of structure fragments (and resulting positional hits) was largely independent of `window_size`, and few unrelated (non-e.8) fragments were captured, indicating that the structure motifs in the surrounding regions were largely limited to the families detected (e.g., e.8.1.4 and e.8.1.6) (Table 2, positions D238 and D328). For positions N297, H398, and H413, at which there was considerable diversity in the SCOP families detected at smaller `window_size` values (<90), omission of these more distantly related positional hits was observed as `window_size` increased. Overall it was observed that `window_size` 90 resulted predominantly in capture of structure fragments from the structure family to which poliovirus RdRp belongs (e.8.1.4). Therefore, at all six positions, `window_size` 90

effectively filtered from among the approximately 160,000 PDB structures only those members of the polymerase/transcriptase families.

Detection of sequence variability in structure context. To determine how sequence variability was distributed among the e.8.1 families, we further categorized the amino-acid variabilities by SCOP family for position N297 at window_size 80 (Table 3). In general approximately 2/3 of the positional hits could be categorized using SCOP; as an example, 320 of the 572 total positional hits were derived from structures that had been classified in SCOP. For this position (at which the greatest sequence variability was observed among the 6 selected), the distribution of amino-acid variability contributed by e.8 superfamily members (e.8.1.1, 2, 4, and 6) was considerably more narrow (5 amino acids (A, F, G, H, N), of which half (80) were N) than that observed overall (13 amino acids). All of the positional hits that coincided in sequence with poliovirus RdRp at N297 were members of the e.8.1.4 (RNA-dependent RNA-polymerase) family, although a minority (5; all of the remaining) of the hits (all RdRp of lambda 3) from this family had H at position 297. Thus, sequence variability was limited to N (94%) and H (6%) within family e.8.1.4, in which poliovirus RdRp is categorized. (This family also includes HCV, FM, lambda 3, BVDV, Norwalk virus, rhinovirus, rabbit hemorrhagic fever virus, and IBVD.) Summarizing the amino-acid occurrences among qualified hits within fold e.8 (comprising 1 superfamily and 4 families), we observed the following amino-acid distributions at position 297 (poliovirus numbering): N=50%, F=26%, G=20%, H=3%, A=0.6%, compared to a much broader distribution of amino-acid variabilities observed within templates outside of fold e.8: V=31%, R=31%, Q=9%, I=6%, H=5%, Y=5%, and C, E, F, K, L, M, and T each \leq 3% (comprising the “tail” in a distribution of sequence

variability). For completeness, amino-acid variabilities derived from positional hits at all six positions examined (see Fig. 3) are summarized in Suppl. Tables 2-7. Distributions of observed amino-acids are shown for all templates and grouped as within or outside of fold 3.8. From this detailed analysis it was possible to categorize the specific amino-acid variabilities per position—thus, StralSV can be used to detect positional trends and anomalies among structures related to the protein of interest.

Effect of window_size on detection of sequence variability. To determine how the window size parameter might affect the detection of amino-acid variability “globally” along the poliovirus RdRp chain, we plotted for each position along the reference structure the detected absolute amino-acid variability versus the number of qualified hits for window_size values 70, 80, and 90 (Fig. 4). As window_size decreased from 90 to 70, there was observed an increase in the number of database structure fragments that contributed positional hits, and a corresponding increase in the amino-acid variabilities: at window_size 70 there were far more (32) positions at which all amino-acids were observed among the positional hits compared to window_size 80 (1 position). At window_size 90 there was considerably less sequence variability detected among the positional hits, with the most variable positions accounting for no more than 14 distinct amino acids observed. The inset in Fig. 4 displays the data points for those positions at which the dominant (most frequently observed) amino-acid occurred at a frequency of at least 80%. Circled data points are those corresponding to window_size 80. There was considerable overlap among window_size values 70, 80, and 90 with respect to positions at which the dominant residue occurred at frequency $\geq 80\%$. Window_size 70

produced 79 positions, 80 produced 74, and 90 produced 79; 62 positions occurred in all three of these data sets. This plot demonstrates that the most dominant amino-acid residues occurred within the same (narrow) positional hit frequency range (0 to <400) regardless of window_size, suggesting that highly conserved positions display relatively little sequence variability regardless of the selected window_size.

Sequence and structure motifs associated with RdRps. Regions corresponding to the well-known sequence motifs characteristic of polymerases were mapped to the positional hit frequency plot for window size 80 (Fig. 5). As mentioned above, regions of high positional hit frequency tended to correspond with helix secondary structure (Fig. 3), but also with defined palm-domain motifs (Fig. 5). Also plotted in Fig. 5 are the frequencies of the dominant residues per position along poliovirus RdRp (lavender plot in figure). Included are residue-position labels for those positions at which the dominant residue frequency exceeded 90% and for which we were able to identify a functional annotation in the literature (see Suppl. Table 1). Although there is no clear criterion for identifying functional residues and short sequence or structure motifs based on StralSV profiles, it was evident that functionally relevant residues tended to emerge when selecting those positions displaying high degrees of conservation in structure context. Furthermore, lowering the window size to just 70 resulted in the capture of many structure fragments with common structure motifs (see last 3 maxima in Fig. 3 graph and Table 2 “other” column for positions H398 and H413), implying that StralSV may enable identification of common structure motifs shared among distantly related proteins.

Discussion

StralSV differs in several respects from other sequence- and structure-based algorithms for comparing proteins. First, by applying an overlapping sliding window to define segments of a structure of interest, the algorithm avoids the pitfalls of algorithms that compare proteins at a global level: by dividing a protein into segments, StralSV enables comparison of structure fragments corresponding approximately in size to super-secondary elements. In this way, portions of a structure can be compared to like fragments in the PDB without the need for explicit splitting of domains. Second, StralSV applies a two-step approach to filtering PDB fragments in order to select those which are most likely to be relevant for a meaningful comparison. For example, the results presented in Tables 2 and 3 and Suppl. Tables 2-7 illustrate how StralSV can be used to quantify and annotate sequence variability among structure fragments that form tight local alignments (determined by span and distance_cutoff parameters), yet have sufficient structure context (determined by window_size) to filter out unrelated fragments. Larger window_size values effectively increase the stringency with which structure fragments are selected, whereas smaller distance_cutoff values also increase this stringency, but more locally, by enforcing tighter local alignments. As seen in Table 2, considerably fewer fragments are selected at window_size 90 than with smaller window_size values for those positions that are sensitive to window_size, but the result set is greatly enriched for fragments that are closely related to the reference structure (i.e., polio RdRp). How much stringency the user wishes to apply in running StralSV may depend on one's research interest and the protein being studied. Smaller

window_size (as well as greater distance_cutoff) parameter values will result in capture of more structure fragments, many of which will be from structures that are more distantly related in terms of their SCOP classification and the taxonomy of the organisms that they are from (Table 4). It must be noted that as the window_size (or distance_cutoff) parameter is relaxed, more “noise” arises in the result set; however, at the same time there is potential for discovering structural relationships and sequence laxity that may not otherwise be discernable when comparing only closely related structures.

Effects of parameter setting on capture of database structure fragments. Not unexpectedly, the frequencies of positional hits tend toward maxima in regions of helical secondary structure, and as window_size is decreased, the numbers of positional hits tend to increase (Figs. 3, 5). Reduction in window_size relaxes constraints on the alignment, and therefore smaller fragments may align more tightly, thereby meeting the distance cutoff. In some regions, the numbers of positional hits do not change significantly with changes in window_size (e.g., peaks up to position 240, and 350-380), indicating that in certain regions there exist conserved structure motifs that are shared only among a specific set of structures. For example, there is observed in the N-terminal regions up to approximately position 140 (Figs. 3, 5) a “polymerase baseline”, which appears to be structurally unique. This region is external to the catalytic tunnel of the polymerase, and may define structural and functional specificity for polio and related viruses. Window_size 90 produced a “polymerase baseline” in the C-terminal region as well, although here the positional hit frequency was highly sensitive to window_size. It may be that only window_size 90 (or greater) effectively filtered out non-polymerase

structure fragments due to the stringency enforced by context, as it appeared that smaller segments of polio RdRp in this region resembled short segments in distantly related proteins.

We observed considerable diversity among the SCOP families represented by fragments detected at different positions at `window_size` values less than 90 (N297, H398, H413). This demonstrated a clear filtering of these positional hits as the `window_size` increased. One must be aware, however, that two parameters (`window_size` and `distance_cutoff`) have an opposing relationship with respect to qualifying a hit: increase in alignment stringency is achieved by increasing `window_size`, but also by decreasing `distance_cutoff`. As `window_size` is decreased, it is necessary to also reduce the `distance_cutoff` in order to avoid an unacceptable increase in the number of “false positive” (considerably less structurally relevant) fragments captured. We achieved a reasonable selection of window-size/distance-cutoff parameter pairings by examining the relationship between these parameters (Fig. 2). The apparent (perhaps unexpected) decrease in the numbers of non-e.8 structure fragments captured by window 50 compared to window 70 at positions N297 and H398 (Table 2 “Other” column) are explained by the high-stringency filtering achieved by the distance cutoff 2.5 Å applied at window 50. Thus, it is desirable to strike a balance between window size and distance cutoff. Small `window_size` values are useful for capturing shorter fragments, but to ensure that the result set is not populated by spurious hits corresponding to ubiquitous secondary structure elements (e.g., alpha helix), one must apply added stringency at the level of `distance_cutoff`. Smaller `window_size` values are appropriate when the user is interested in focussing the

analysis on a relatively small structure motif, which may occur with or without the surrounding structural context in the reference structure. In capturing these smaller database structure fragments, one is cautioned to enforce tighter alignments in order to assure that the resulting qualified hits are relevant to the study.

Detection of Sequence Motifs. Regions corresponding to the well-known sequence motifs A-G, characteristic of polymerases, were mapped to the positional hit frequency plot for window size 80 (Fig. 5). All categories of polymerase (RdRp, RdDp/RT, DpRp, DpDp) are recognized to contain sequence motifs A-D. Identification of motifs E, F, and G, however, has been somewhat obscured by the greater diversity among sequences assigned to these structure motifs. O'Reilley (1998) reported motif E as being exclusive to RdDps and RdRps, and motifs F (Bruenn 2003, Xu 2003) and G (Xu 2003) were identified in poliovirus RdRp. Motif F was identified in phi6 (Butcher 2001), BVDV and CHV (Choi 2004), reovirus, norovirus, and rhinovirus (Pan 2007), and RHDV and HIV1 (Ferrer-Orta 2004), and reovirus, phi6, BVDV, HCV, norovirus, and HIV (Pan 2007). A detailed structure-based comparison of these polymerases using StralSV may clarify the assignment of motifs A-G among the polymerase classes.

Detection of sequence variability in structure context. In examining the amino-acid variability versus qualified hit frequency (Fig. 4 inset, circled data points) for 11 highly conserved positions (Fig. 5 dots) that had been functionally annotated (Suppl. Table 1), it appeared that the numbers of positional hits and the degree of amino-acid variability observed for high-frequency positions was largely independent of window_size. This implies (at least for the positions that were examined in this study) that the functionally relevant positions are consistently detected as high-frequency-residue positions

regardless of `window_size`. Therefore, the StralSV algorithm is sufficiently sensitive to detect structurally or functionally conserved residues even when the parameters may not be perfectly tuned.

StralSV is especially useful for identifying highly conserved residues at positions that occur in regions in which there are large numbers of positional hits; dominant residue frequency cannot be considered significant in regions of structural conservation (e.g., the N-terminal region up to about position 140 in poliovirus RdRp), whereas identification of dominant residues occurring with high frequency at positions with large numbers of positional hits may help identify residues at positions that are structurally and/or functionally significant. For example, more than 250 positional hits contributed to residue counts at positions D233 and D328 (`window_size` 80, Fig. 5), known to be critical for RdRp function. Aspartic acid occurred at close to 100% frequency at these positions. The structure fragments contributing to these data points comprised RdRps and dsRNA phage polymerases (SCOP family e.8.1.4). The combination of many positional hits and low amino acid variability may provide a means of identifying key functional residues

Results from StralSV analysis can be used to characterize residue positions in a reference protein by detection of similar locations in other proteins (sometimes from quite distant organisms and different assigned structure classifications), in which corresponding residue positions within similar structural motifs are observed. Such analyses may potentiate rapid identification of invariant (as well as unusual or unexpected) residues, which in many cases are essential to a protein's function. It also may enlighten studies of newly discovered natural or engineered mutations that have

not yet been observed in the sequence databases. Results from StralSV structure similarity searches performed against large sets of structurally related proteins can facilitate refinement of constructed homology models by suggesting corrections to the query-template alignments (using an approach similar to that of Chodanowski 2008) or by providing a list of possible conformational variants of corresponding structural fragments for loop-building procedures. Calculated residue-residue correspondences can be used to evaluate pure sequence alignment methods and also to derive structural environment-specific substitution matrices, which have been shown to be useful for detection of remote homologs (Gelly 2005). When applied to experimentally solved structures, StralSV also could facilitate identification of structural motifs (local conformations) that have not yet been observed in PDB. Such findings could aid in the discovery of previously unidentified structural motifs or suggest refinement of constructed structural models in particular regions.

Conclusions

StralSV is a new algorithm for detecting closely related structure fragments from a structure database (PDB or user-defined) and quantifying residue frequency from tight local structure alignments. Input parameters to StralSV can be varied in order to adjust the stringency with which structure fragments are selected (via `window_size` and `LGA_S` cutoff) or with which local alignments are made (via `distance_cutoff`, `span_size`, and minimum span length), thereby providing the user with flexibility in detecting similar structure fragments. High-stringency parameter settings will effectively filter out all but highly structurally similar fragments and will impose very tight local alignments, whereas

low-stringency parameters will enable detection of more distantly related structures, which may be of interest, for example, when the user wishes to detect distant evolutionary relationships among proteins or to test the range of possible sequence variability that might be expected to be tolerated within a given structure motif. It should be emphasized, however, that StralSV safeguards against degradation of sequence variability data quality by enforcing structure context upon local alignments in a two-step process of identifying “qualified” hits.

It has long been recognized that peptides with very different sequences may have similar tertiary structures. In this work we applied StralSV in a study of the RNA-dependent RNA polymerase of poliovirus and demonstrated that the algorithm could be used to determine regions of the protein that were relatively unique (the N-terminal region) or that shared structural similarity (e.g., C-terminal motifs) with structures that were distantly related (non-e.8 SCOP classifications), and that by quantifying residue frequencies among many (hundreds or even thousands) of residue-residue pairs extracted from local alignments, one can infer potential structural or functional importance of specific residues that are determined to be highly conserved or that deviate from a consensus. We further demonstrated that considerable detailed structural and phylogenetic information can be derived from StralSV profiles.

StralSV is available as a web service at <http://as2ts.llnl.gov/stralsv/>.

List of Abbreviations

RdRp—RNA-dependent RNA polymerase

StralSV—Structure-alignment-based sequence variability

SCOP—Structure Classification of Proteins database

PDB—Protein Databank database

Authors' contributions

Acknowledgements and Funding

This work was conducted at Lawrence Livermore National Laboratory under DOE Contract DE-AC52-07NA27344. The work was supported by an LLNL-LLNS internally funded grant to TV and CZ through the Laboratory Directed Research and Development program. TV was also partially supported by the National Science Foundation, through the Independent Research and Development program. Any opinion, finding, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

- Bruenn JA. 2003. A structural and primary sequence comparison of the viral RNA-dependent RNA polymerases. *Nucleic Acids Research* 31:1821-1829.
- Chodanowski P, A Grosdidier, E Feytmans, O Michielin. 2008. Local Alignment Refinement Using Structural Assessment. *PLoS ONE* 3(7): e2645. doi:10.1371.
- Choi KH, JM Groarke, DC Young, RJ Kuhn, JL Smith, DC Pevear, MG Rossmann. 2004. The structure of the RNA-dependent RNA polymerase from bovine viral diarrhea virus establishes the role of GTP in de novo initiation. *PNAS* 101:4425-4430.
- Domingues FS, J Rahnenführer, T Lengauer. 2004. Automated clustering of ensembles of alternative models in protein structure databases. *Protein Eng Des Sel.* 17(6):537-43.
- Dundas J, TA Binkowski, B DasGupta, J Liang. 2007. Topology independent protein structural alignment. *BMC Bioinformatics.* 8:388.
- Ferrer-Orta C, A Arias, R Perez-Luque, C Escarmis, E Domingo, N Verdaguer. 2004. Structure of foot-and-mouth disease virus RNA-dependent RNA polymerase and its complex with a template-primer RNA. *Journal of Biological Chemistry* 279: 47212-47221.
- Gelly JC, L Chiche, J Gracy 2005. EvDTree: structure-dependent substitution profiles based on decision tree classification of 3D environments. *BMC Bioinformatics.* 6:4.
- Godzik A. 1996. The structural alignment between two proteins: is there a unique answer? *Protein Science.* 5:1325–1338.
- Gohara DW, S Crotty, JJ Arnold, JD Yoder, R Andino, CE Cameron. 2000. Poliovirus RNA-dependent RNA polymerase (3Dpol): Structural, biochemical, and biological analysis of conserved structural motifs A and B. *Journal of Biological Chemistry* 275:25523-25532.
- Hansen JL, AM Long, SC Schultz. 1997. Structure of the RNA-dependent RNA polymerase of poliovirus. *Current Biology* 5:1109-1122.
- Hobson SD, ES Rosenblum, OC Richards, K Richmond, K Kirkegaard, SC Schultz. 2001. Oligomeric structures of poliovirus polymerase are important for function. *EMBO Journal* 20:1153-1163.
- Kim C, CH Tai, B Lee. 2009. Iterative refinement of structure-based sequence alignments by Seed Extension. *BMC Bioinformatics.* 10:210.
- Kolbeck B, P May, T Schmidt-Goenner, T Steinke, EW Knapp. 2006. Connectivity independent protein-structure alignment: a hierarchical approach. *BMC Bioinformatics.* 7:510.

Lesburg CA, MB Cable, E Ferrari, Z Hong, AF Mannarino, PC Weber. 1999. Crystal structure of the RNA-dependent RNA polymerase from hepatitis C virus reveals a fully encircled active site. *Nature Structural Biology* 6:937-943.

Lindqvist Y, G Schneider. 1997. Circular permutations of natural protein sequences: structural evidence. *Curr Opin Struct Biol.* 7(3):422-427.

Marcotte LL, AB Wass, DW Gohara, HB Pathak, JJ Arnold, DJ Filman, CE Cameron, JM Hogle. 2007. Crystal structure of poliovirus 3CD protein: virally encoded protease and precursor to the RNA-dependent RNA polymerase. *Journal of Virology* 81:3583-3596.

Mayr G, FS Domingues, P Lackner. 2007. Comparative analysis of protein structure alignments. *BMC Structural Biology* 7:50.

Menke M, B Berger, L Cowen. 2008. Matt: local flexibility aids protein multiple structure alignment. *PLoS Comput Biol.* 4(1).

Milik M, S Szalma, K Olszewski. 2003. Common Structural Cliques: a tool for protein structure and function analysis. *Protein Engineering* 16(8):543-552.

Murzin, AG, SE Brenner, T Hubbard, C Chothia. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology* 247: 536-540.

O'Reilly EK, CC Kao. 1998. Analysis of RNA-dependent RNA polymerase structure and function as guided by known polymerase structures and computer predictions of secondary structure. *Virology* 252:287-303.

Pan J, VN Vakharia, YJ Tao. 2004. The structure of a birnavirus polymerase reveals a distinct active site topology. *PNAS* 104:7385-7390.

Pfeiffer JK, K Kirkegaard. 2003. A single mutation in poliovirus RNA-dependent RNA polymerase confers resistance to mutagenic nucleotide analogs via increased fidelity. *PNAS* 100:7289-7294.

Rost B. 1999. Twilight zone of protein sequence alignments. *Protein Engineering.* **12** (2), 85-94.

Rost B and Sander C. 1993. Prediction of secondary structure at better than 70% accuracy. *Journal of Molecular Biology* 232: 584-599.

Salem S, MJ Zaki, C Bystroff. 2010. FlexSnap: Flexible Non-sequential Protein Structure Alignment. *Algorithms for Molecular Biology*, 5:12.

Shatsky M, R Nussinov, H Wolfson. 2002. Flexible protein alignment and hinge detection. *Proteins: Structure, Function, and Bioinformatics* 48:242-256.

Xu X, Y Liu, S Weiss, E Arnold, SG Sarafinos, J Ding. 2003. Molecular model of SARS coronavirus polymerase: implications for biochemical functions and drug design. *Nucleic Acids Research* 31:7117-7130.

Yakunin AF, AA Yee, A Savchenko, AM Edwards, CH Arrowsmith. 2004. Structural proteomics: a tool for genome annotation. *Curr Opin Chem Biol.* 2004;8:42–8.

Ye Y, Godzik A. 2003. Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics* 19, Suppl. 2:ii246-255.

Zemla A. 2003. LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids research* 31:3370-3374.

Zemla A, B Geisbrecht, J Smith, M Lam, B Kirkpatrick, M Wagner, T Slezak, CE Zhou. 2007. STRALCP--structure alignment-based clustering of proteins. *Nucleic Acids Research* doi:10.1093/nar/gkm1049.

Figures

Fig. 1. Overview of process and data types used in the StralSV algorithm.

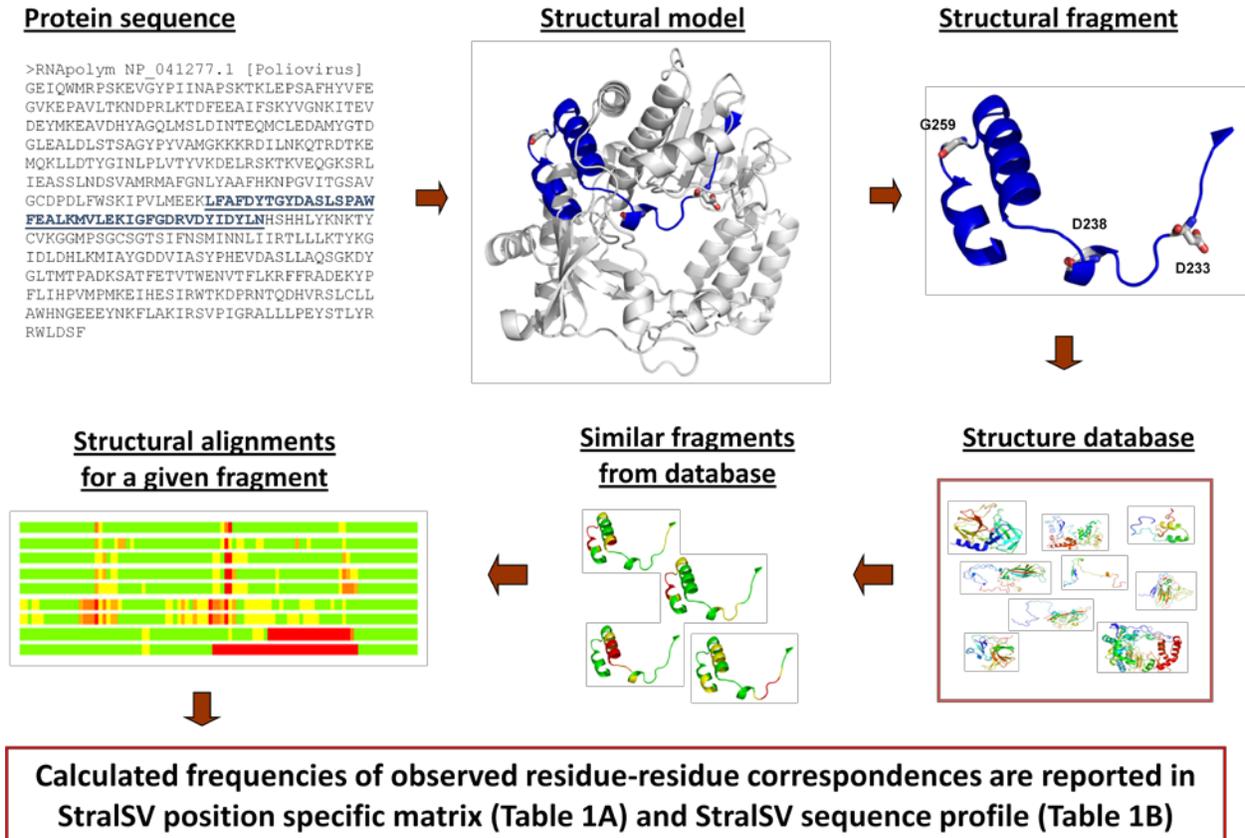


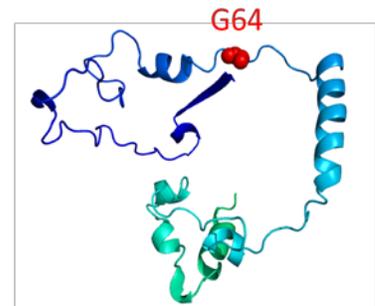
Fig. 2. Effect of window size and distance cutoff parameter combinations on capture of structure fragments. A: sequence variability profile around position G64 generated by running StralSV against the PDB database; red type: data for position G64. B: Numbers of structure fragments selected from a custom library comprising 38 polymerase chains plus randomly selected structures from PDB; red type: numbers corresponding to distance cutoff and window_size values selected as defaults for StralSV web service. C: Structure fragment of poliovirus polymerase (1rdr) corresponding to the N-terminal region containing position G64; color variations dark blue to bright turquoise indicate N-to-C-terminal direction of chain.

Sequence variability profile (around G64) based on the complete PDB (147,634 chains)

#	AA	ord	d/+	2nd/+	3rd/+	+	variety
LSV_P:	K	61	66.7	28.1	5.3	57	KPV
LSV_P:	Y	62	73.7	26.3	0.0	38	YH
LSV_P:	V	63	57.9	34.2	7.9	38	KVI
LSV_P:	G	64	97.4	2.6	0.0	38	GS
LSV_P:	N	65	70.0	25.0	5.0	40	NDP
LSV_P:	K	66	47.6	40.5	11.9	42	TKV
LSV_P:	I	67	33.3	25.6	23.1	39	IKELN

Number of similar structural motifs identified in the dataset of 1460 selected chains

6.5	137	112	88	51	178	42	50	41	40	49	38
6.0	121	212	71	53	117	45	47	41	43	46	38
5.5	123	137	81	47	82	39	44	39	40	40	38
5.0	115	134	53	48	59	40	38	39	40	40	38
4.5	83	67	46	54	58	39	38	39	38	40	38
4.0	73	58	42	43	41	38	38	38	39	38	38
3.5	58	49	38	38	41	38	38	38	38	39	38
3.0	48	40	40	39	38						
2.5	44	38	38	38							
2.0	39	34	33	35	32	38	34	34	33	33	33
1.5	38	36	32	31	29	31	28	27	28	23	20
1.0	28	24	22	15	13	15	13	13	14	13	13
d/w	20w	30w	40w	50w	60w	70w	80w	90w	100w	120w	150w



N-terminal region: 1 - 150
Unique for most (38) RDRPs deposited in PDB

Fig. 3. Effect of window size on numbers of positional hits per position. Red: window size 70; blue: 80; green: 90. Vertical lines with labels: positions at which SCOP and EC identifiers were quantified for templates contributing to positional variability data (see Table 2 and Suppl. Tables 2-7). Along x-axis: secondary structure predictions (see Methods). Inset: structure model of poliovirus polymerase upon which have been labeled the six positions highlighted in main figure.

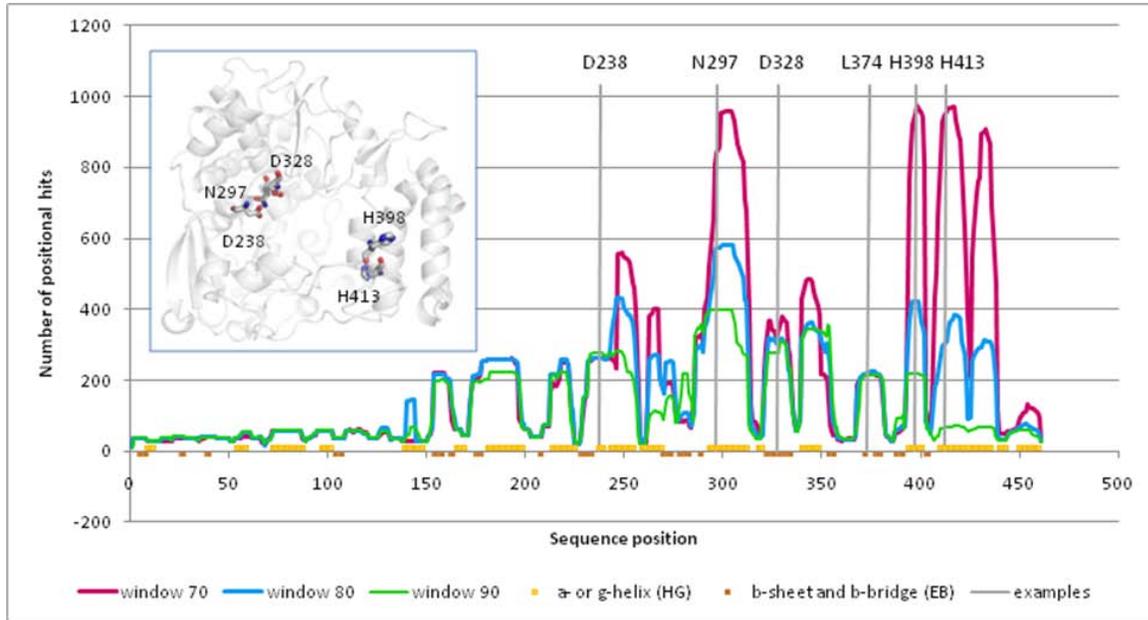


Fig. 4. Amino-acid variability (y-axis) versus number of positional hits (x-axis) for three window sizes. Red circles: window size 70; blue diamonds: 80; green triangles: 90. Inset: positional hits at which the dominant residue occurs at frequency $\geq 80\%$. Circles: positional hit, variability coordinate pairs corresponding to the 11 positions shown in Fig. 5).

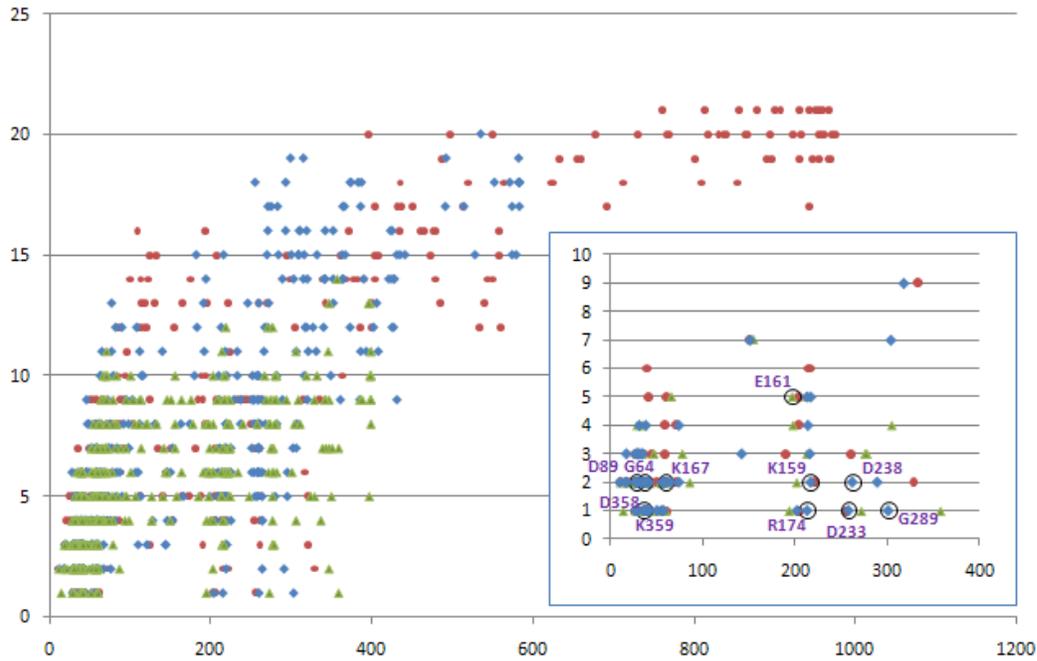


Fig. 5. Frequencies of dominant residues and correspondence of positional hit frequency with polymerase sequence motifs A-G. Lavender plot: frequencies (quantified along right y-axis) of dominant residue per sequence position along poliovirus polymerase chain; blue plot: positional hit frequency at window size 80; labeled dots: high-frequency residues that have been functionally annotated (see Suppl. Table 1).

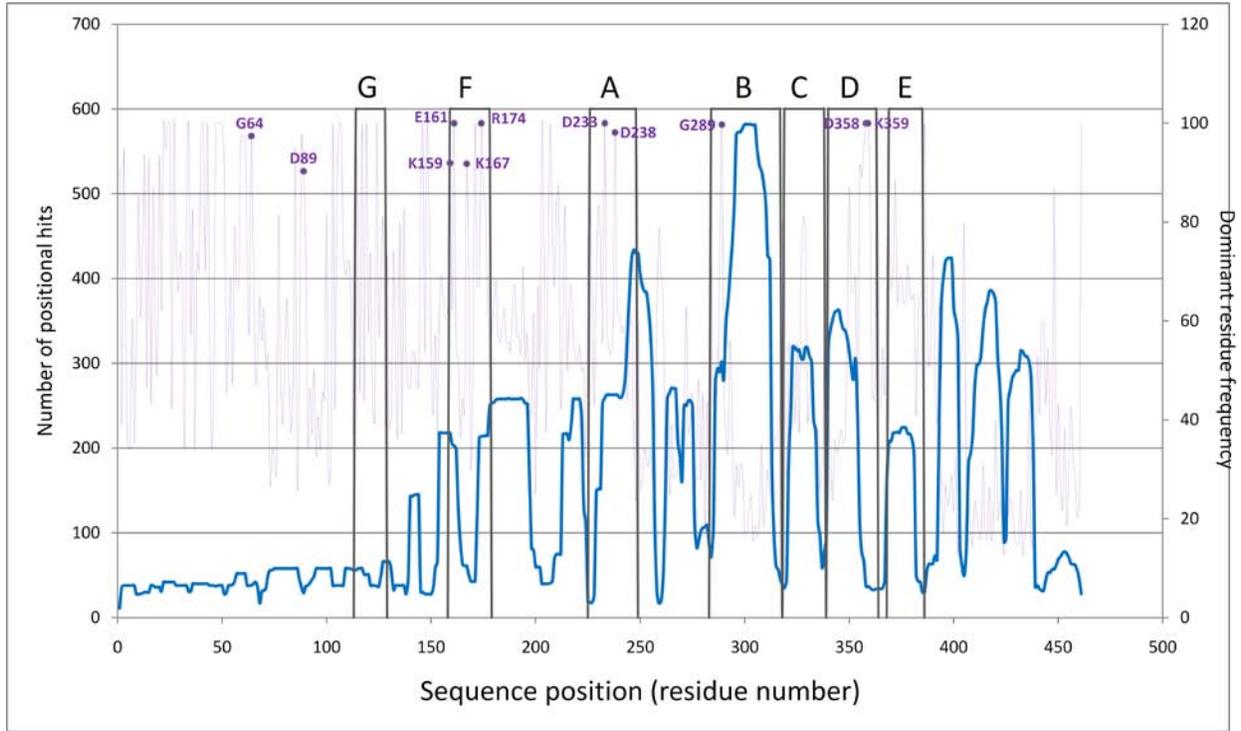


Table 1A. Excerpt of a sample StralSV output “matrix” file for positions 229 through 269 of poliovirus RNA-dependent RNA polymerase (RdRp) analysis. The heading provides the total number of templates searched plus other run parameters. The first column contains the amino acid of the query sequence at the position indicated in the second column. Each row in the table comprises a tally of the amino acids of each type (including non-standard indicated by ‘X’) from the corresponding positions in all qualified template fragments (positional hits).

```

Number of all PDB chains (SEQRES.all_list.11_03_09): 147,634
Structure: Polio_RNAPolym Length: 461 MIN_span: 5 Structure_similarity: 55 LGA_distance: 4.0 Window: 90
Legend: 's' AA reference sequence, 'd' dominant amino acid, '+' number of hits per position ( max '+' : n = 398 )

```

AA	Rname	A	V	L	I	P	M	F	W	G	S	T	C	Y	N	Q	D	E	K	R	H	X
L	229	4	0	21	10	0	0	1	0	103	0	0	0	0	0	0	0	0	0	0	4	0
F	230	0	4	4	3	0	8	127	0	0	0	0	5	0	0	0	0	0	0	0	0	4
A	231	30	0	0	0	0	0	0	0	116	0	2	5	0	0	4	0	0	0	0	0	0
F	232	8	10	7	4	0	0	34	0	0	0	49	0	140	0	0	5	0	0	0	0	0
D	233	0	0	0	0	0	0	0	0	0	0	0	0	0	0	272	0	0	0	0	0	0
Y	234	0	49	4	11	0	0	0	0	0	152	0	60	0	0	0	0	0	0	0	0	0
T	235	5	0	0	0	0	0	0	0	91	25	0	0	0	0	0	4	7	144	0	0	0
G	236	23	0	0	0	0	0	0	23	0	0	139	0	13	5	48	0	11	13	0	1	0
Y	237	4	0	0	0	0	0	152	37	4	0	0	7	28	0	0	0	0	0	44	0	0
D	238	0	0	0	0	0	0	0	0	0	0	0	4	0	0	268	4	0	0	0	0	0
A	239	41	0	0	0	0	0	0	14	165	56	0	0	0	0	0	0	0	0	0	0	0
S	240	0	0	0	0	0	44	0	0	37	160	0	0	18	7	0	0	0	5	0	0	0
L	241	0	151	20	22	0	4	0	44	0	0	4	0	0	16	0	0	0	0	10	0	0
S	242	0	0	0	0	44	0	0	0	24	165	10	0	3	13	0	0	4	0	10	0	0
P	243	0	0	0	0	37	0	0	7	44	17	0	0	0	10	0	142	0	20	0	0	0
A	244	34	7	0	8	10	0	0	44	0	14	0	4	0	114	4	12	0	6	13	0	0
W	245	10	20	48	0	0	10	30	0	0	0	0	0	0	0	152	0	0	4	0	0	0
F	246	0	4	32	140	0	14	30	0	0	3	4	0	0	0	0	0	0	54	0	0	0
E	247	19	2	0	0	0	0	0	0	7	0	0	0	10	12	44	18	22	145	2	0	0
A	248	44	135	59	10	0	0	0	10	0	12	5	0	5	0	0	0	0	1	0	0	0
L	249	20	0	30	72	0	11	0	0	3	0	0	0	0	0	145	0	0	0	0	0	0
K	250	3	0	16	14	0	10	10	0	15	0	0	29	0	5	0	144	20	0	0	0	0
M	251	11	0	0	0	0	31	0	0	0	144	0	3	0	0	22	20	4	0	0	4	0
V	252	8	27	2	168	0	0	0	9	0	0	0	3	0	0	0	6	0	0	0	0	0
L	253	0	0	52	10	0	14	0	0	0	0	0	140	0	7	0	0	0	0	0	0	0
E	254	3	13	9	0	0	0	0	0	2	0	1	0	140	16	23	7	0	1	0	0	0
K	255	12	0	0	2	0	0	0	0	0	4	128	7	10	0	2	1	33	3	0	0	0
I	256	0	0	17	14	0	0	16	0	1	0	85	7	0	0	0	0	1	0	0	0	0
G	257	0	0	0	0	0	0	0	20	23	0	0	7	0	0	1	0	0	0	0	0	0
F	258	0	0	0	0	0	18	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0
G	259	0	0	0	0	0	0	0	14	0	0	0	0	0	0	0	0	0	0	0	0	0
D	260	2	0	0	0	0	0	0	0	0	0	0	0	2	0	14	0	0	0	0	0	0
R	261	0	0	0	0	2	0	0	2	0	0	0	0	0	0	0	0	14	0	0	0	0
V	262	0	46	0	2	5	0	0	0	7	3	0	0	0	13	0	0	4	2	0	0	0
D	263	2	13	0	0	0	0	0	10	0	2	5	3	0	7	12	14	0	32	0	0	0
Y	264	4	13	44	12	0	0	3	0	0	10	0	17	0	0	0	0	0	0	2	0	0
I	265	13	0	12	37	2	0	32	0	0	7	0	0	0	10	0	0	0	0	0	0	0
D	266	2	3	0	10	0	0	0	0	2	0	0	0	8	2	26	50	10	0	0	0	0
Y	267	3	0	4	0	0	0	0	0	2	42	0	17	10	0	16	0	0	10	7	0	0
L	268	0	2	63	2	0	5	0	0	31	0	0	2	0	0	0	0	0	0	3	2	0
N	269	0	13	38	3	0	10	0	0	3	6	15	0	13	2	0	0	0	0	0	0	0

Table 1B. Excerpt of a sample StralSV output “profile” file for positions 229 through 269 of poliovirus polymerase. #: file name; AA: query amino acid; ord: sequence position number; Rname: PDB residue name; rnk: rank (position of the query amino acid in the “variety” list); s/+: percent of all residues corresponding to query; d/+: percent of dominant residue; 2nd/+: percent of second most dominant residue; 3rd/+: percent of third most dominant residue; +/n: percent of total templates selected in given experiment contributing to data at a given position; +: total number of templates contributing to data at a given position; nv: total number of amino acid variants; variety: list of variants in order of frequency.

#	AA	ord	Rname	rnk	s/+	d/+	2nd/+	3rd/+	+/n	+	nv	variety
LSV_P:	L	229	229	2	14.7	72.0	14.7	7.0	35.9	143	6	GLIAHF
LSV_P:	F	230	230	1	81.9	81.9	5.2	3.2	38.9	155	7	FMYVLXI
LSV_P:	A	231	231	2	19.1	73.9	19.1	3.2	39.4	157	5	SAYDC
LSV_P:	F	232	232	3	13.2	54.5	19.1	13.2	64.6	257	8	YTFVALDI
LSV_P:	D	233	233	1	100.0	100.0	0.0	0.0	68.3	272	1	D
LSV_P:	Y	234	234	2	21.7	55.1	21.7	17.8	69.3	276	5	TYVIL
LSV_P:	T	235	235	3	9.1	52.2	33.0	9.1	69.3	276	6	RSTKAE
LSV_P:	G	236	236	4	8.3	50.4	17.4	8.3	69.3	276	9	CDAGNRKQX
LSV_P:	Y	237	237	4	10.1	55.1	15.9	13.4	69.3	276	7	FHWYCAG
LSV_P:	D	238	238	1	97.1	97.1	1.4	1.4	69.3	276	3	DYE
LSV_P:	A	239	239	3	14.9	59.8	20.3	14.9	69.3	276	4	STAG
LSV_P:	S	240	240	3	13.7	59.0	16.2	13.7	68.1	271	6	TFSNQR
LSV_P:	L	241	241	4	7.4	55.7	16.2	8.1	68.1	271	8	VWILQHMC
LSV_P:	S	242	242	3	8.8	60.4	16.1	8.8	68.6	273	8	TPSQCHKN
LSV_P:	P	243	243	3	13.4	51.3	15.9	13.4	69.6	277	7	EGPRSQW
LSV_P:	A	244	244	3	12.6	42.2	16.3	12.6	67.8	270	12	NWASRDPIVKCQ
LSV_P:	W	245	245	3	10.9	55.5	17.5	10.9	68.8	274	7	DLWVAFH
LSV_P:	F	246	246	4	10.7	49.8	19.2	11.4	70.6	281	8	IRLFMVCT
LSV_P:	E	247	247	5	6.4	51.6	15.7	7.8	70.6	281	10	RDKAEQNSVH
LSV_P:	A	248	248	3	15.7	48.0	21.0	15.7	70.6	281	9	VLATIGCNH
LSV_P:	L	249	249	3	10.7	51.6	25.6	10.7	70.6	281	6	EILAMS
LSV_P:	K	250	250	3	7.5	54.1	10.9	7.5	66.8	266	10	ECKLGLIMFQA
LSV_P:	M	251	251	2	13.0	60.3	13.0	9.2	60.1	239	8	SMDEAKXY
LSV_P:	V	252	252	2	12.1	75.3	12.1	4.0	56.0	223	7	IVGAEYL
LSV_P:	L	253	253	2	23.3	62.8	23.3	6.3	56.0	223	5	YLMIQ
LSV_P:	E	254	254	2	10.7	65.1	10.7	7.4	54.0	215	10	QEDVLKATYH
LSV_P:	K	255	255	2	16.3	63.4	16.3	5.9	50.8	202	10	CKANYTRIDE
LSV_P:	I	256	256	4	9.9	60.3	12.1	11.3	35.4	141	7	CLFIYGR
LSV_P:	G	257	257	2	39.2	45.1	39.2	13.7	12.8	51	4	SGYD
LSV_P:	F	258	258	1	85.7	85.7	14.3	0.0	5.3	21	2	FY
LSV_P:	G	259	259	1	100.0	100.0	0.0	0.0	3.5	14	1	G
LSV_P:	D	260	260	1	77.8	77.8	11.1	11.1	4.5	18	3	DAN
LSV_P:	R	261	261	1	77.8	77.8	11.1	11.1	4.5	18	3	RMG
LSV_P:	V	262	262	1	56.1	56.1	15.9	8.5	20.6	82	8	VQSPKTIR
LSV_P:	D	263	263	2	14.0	32.0	14.0	13.0	25.1	100	10	KDVQWNTCAS
LSV_P:	Y	264	264	2	16.2	41.9	16.2	12.4	26.4	105	8	LYVITAFH
LSV_P:	I	265	265	1	32.7	32.7	28.3	11.5	28.4	113	7	IFALQTP
LSV_P:	D	266	266	2	23.0	44.2	23.0	8.8	28.4	113	9	EDIKNVASQ
LSV_P:	Y	267	267	2	15.3	37.8	15.3	14.4	27.9	111	9	TYDNRHLAS
LSV_P:	L	268	268	1	57.3	57.3	28.2	4.5	27.6	110	8	LSMHVIYX
LSV_P:	N	269	269	4	12.6	36.9	14.6	12.6	25.9	103	9	LCVNMTISQ

Table 2. SCOP categorization of template fragments selected at each of six positions along poliovirus polymerase. Position: residue and position number in poliovirus polymerase; Window: window_size parameter settings (w: window); RdRp e.8.1.4: RNA-dependent RNA polymerase; DsPhage e.8.1.6: dsRNA phage RNA-dependent RNA-polymerase; RT e.8.1.2: Reverse transcriptase; DNA pol1 e.8.1.1: DNA polymerase I; Other not e.8: all other SCOP families.

Position	Window	DNA				
		RdRp e.8.1.4	DsPhage e.8.1.6	RT e.8.1.2	pol1 e.8.1.1	Other not e.8
D238	w90	85	32	0	0	3
	w80	85	32	0	0	0
	w70	79	32	0	0	0
	w50	79	0	0	0	0
N297	w90	85	32	78	0	9
	w80	85	32	41	2	160
	w70	85	32	71	3	289
	w50	78	32	0	4	197
D328	w90	84	32	3	0	6
	w80	84	32	8	0	17
	w70	84	32	4	0	34
	w50	71	2	0	0	3
L374	w90	77	0	0	0	2
	w80	80	0	0	0	3
	w70	82	0	1	0	3
	w50	59	0	0	0	2
H398	w90	79	0	0	0	3
	w80	80	0	0	0	48
	w70	68	0	0	2	376
	w50	36	0	0	0	182
H413	w90	37	0	0	0	1
	w80	45	0	0	0	54
	w70	40	0	0	2	392
	w50	37	0	0	13	695

Table. 3. Effect of window_size on sequence variability in SCOP-categorized structure fragments at position N297.

Position: 297							
Window: 80							
Categories: 33							
PDB templates: 399							
SCOP_IDs: 320							
Templates	all PDB templates	e.8 templates	RNA-dependent RNA polymerase	dsRNA phage RNA-dependent RNA polymerase	DNA polymerase I	Reverse transcriptase	templates not e.8
Scop ID		e.8	e.8.1.4	e.8.1.6	e.8.1.1	e.8.1.2	
max LGA		100.00	100.00	76.94	57.01	59.17	76.83
ag LGA		75.42	85.15	75.13	56.96	56.38	57.79
avg ID		27.31	37.28	13.42	3.74	18.63	6.24
matches		51	61	40	31	39	35
AA	572	160	85	32	2	41	160
A	1	1			1		
C	8						3
D	1						
E	3						1
F	67	42			1	41 ¹	5
G	44	32		32 ²			
H	14	5	5 ³				8
I	19						9
K	2						2
L	11						5
M	2						2
N	224	80	80 ⁴				
P							
Q	15						14
R	82						49
S							
T	15						4
V	54						50
W							
Y	10						8
X							
¹ F297 (N=41): reverse transcriptase of HIV							
² G297 (N=32): dsRNA phage RNA-dependent RNA-polymerase of phi6							
³ H297 (N=5): RdRp of lambda 3							
⁴ N297 (N=80): RdRp proteins from HCV (21), FM (8), lambda3 (5), polio (6), BVDV (3), Norwalk virus (3), rhinovirus (3), rabbit hemorrhagic fever virus (2), IBDV (1)							

Table 4. Taxonomic diversity represented by positional hits detected at position N297.

window	50	70	80	90
No. of PDBs	189	198	199	141
ARCHAEA	+	+	+	
BACTERIA	+	+	+	+
BIRD	+			
FISH			+	
FUNGUS	+			
INSECT	+	+	+	
MAMMAL	+	+	+	+
PHAGE	+		+	+
VIRUS	+	+	+	+
WORM	+	+		
YEAST	+	+	+	

Supplementary Figures and Tables

Suppl. Fig. 1. Illustration of a span. A span is a contiguous set of residues from two structures comprising a tight local alignment. Three spans are identified as regions A1:A2, B1:B2, and C1:C2. Short line segments at A2, B2, C1, and C2 clarify residues at which a span begins or ends. Color ranging from medium blue through yellow specify chain direction from the N- to the C-terminus. Only residue-residue correspondences that occur within identified spans (i.e., that meet the specified criteria) contribute to the amino-acid counts in the matrix and profile output files.

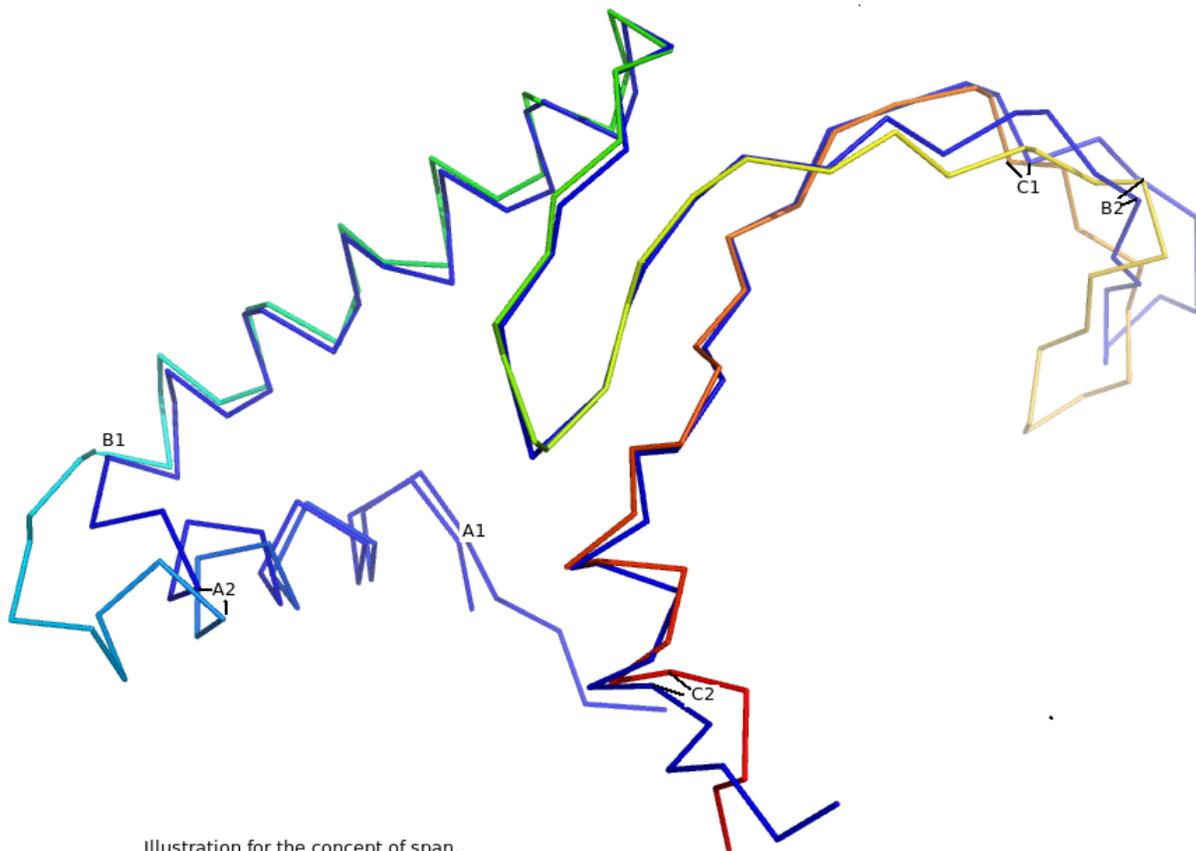


Illustration for the concept of span.
Three spans (A1:A2), (B1:B2), (C1:C2) have been identified
for these two aligned 2D protein structures.

Suppl. Table 1. Dominant (frequency >90%) residues and functional annotations (when known) for StralSV analysis of poliovirus RdRp.

Position	Dominant Residue	Polio Residue	Structural and Functional Annotations	Literature Citations
3	I			
13	G			
15	P			
20	P			
22	K			
23	T			
24	K			
25	L			
27	P			
34	F		anchors N-terminal strand; interface II residue	Hobson 2001
40	P			
41	A			
43	L			
47	D		essential for viability; functional (?)	O'Reilley 1998, Diamond 1994
48	P			
49	R		essential for viability; functional (?)	O'Reilley 1998, Diamond 1994
50	L			
51	K		mutation results in temperature sensitivity; mutation prevents viral RNA accumulation	O'Reilley 1998, Diamond 1994; Marcotte 2007
59	F			
60	S			
61	K		essential for viability	O'Reilley 1998; Xu 2003; Richards 1996
64	G		fidelity; ribavirin resistance	Pfeiffer & Kirkegaard 2003; Marcotte 2007; Gohara 2000
79	K	D	domain-interface residue	Diamond 1994; Marcotte 2007
80	E	H	domain-interface residue	Diamond 1994; Marcotte 2007
85	L			
88	L			
89	D		binds divalent cation	Diamond 1994; Marcotte 2007; O'Reilley 1998
103	G			
106	G			
107	L			
117	G			
119	P			
124	G			
146	D			
147	Y		essential for viability	O'Reilley 1998
158	K		functional (?)	Xu 2003
161	E		<see refs>	Lesburg 1999; Xu 2003
167	K		contacts phosphate of incoming nucleotide	Gohara 2000
171	G			
174	R		contacts phosphate of incoming nucleotide	Bruenn 2003; Xu 2003; Gohara 2000
203	G			
207	G			
233	D		coordinates catalytically essential metals; binds metal ion; active site residue	O'Reilley 1998; Hansen 1997, Xu 2003, Genbank; Burns 1989
238	D		discriminates d/NTPs; hydrogen bonds with N297; involved in nucleotide selection: couples binding of nucleotide with correct sugar configuration; active site residue	O'Reilley 1998; Hansen 1997; Xu 2003; Genbank; Gohara 2000
289	G		putatively binds to nucleic acid; active site residue	O'Reilley 1998; Xu 2003; Genbank;
355	T			
357	A			
358	D		mutation results in loss of uridylylation	Marcotte 2007
359	K		mutation results in loss of uridylylation	O'Reilley 1998; Marcotte 2007
386	F			
461	F			

Suppl. Table 2. Effect of window_size on selection of fragments and sequence variability in SCOP-categorized structure fragments at position D238. Column/row labels: PDB: all templates selected by StralSV from PDB; e.8: templates selected from PDB with SCOP classification e.8; not e.8: templates selected from PDB with SCOP identity that is other than e.8; avg LGA_S: average of LGA_S scores for indicated templates; avg seq_ID: average of sequence identity scores for indicated templates; avg R-Rs/template: average of the number of residue-residue matches per template; # PDB/SCOP_IDs: the number of templates selected by StralSV for each type of template (PDB or SCOP); # SCOP categories: the number of different SCOP categories of all “e.8” and all “not e.8” templates.

Position: D238												
window size	window 90			window 80			window 70			window 50		
classification	PDB	e.8 ¹	not e.8	PDB	e.8	not e.8	PDB	e.8	not e.8	PDB	e.8	not e.8
avg LGA_S		82.11	70.58		76.16			80.05			88.74	
avg seq_ID		25.88	14.1		23.04			22.26			29.36	
avg R-Rs/template		63	44		48			43			32	
# PDB/SCOP_IDs	275	117	1	261	113	0	260	111	0	207	79	0
# SCOP categories		2	1		2	0		3	0		1	0
AA: (below)												
A												
C												
D	267	116	1	256	112		255	111		207	79	
E	4	1		5	1		2					
F												
G												
H												
I												
K												
L												
M												
N												
P												
Q												
R												
S												
T												
V												
W												
Y	4						3					
X												
¹ Types of e.8 templates included in e.8 of this study: e.8.1.1, e.8.1.2, e.8.1.4, e.8.1.6												

Suppl. Table 3. Effect of window_size on sequence variability in SCOP-categorized structure fragments at position N297. For column headings see Suppl. Table 2.

Position: N297												
window size	window 90			window 80			window 70			window 50		
classification	PDB	e.8	not e.8	PDB	e.8	not e.8	PDB	e.8	not e.8	PDB	e.8	not e.8
avg LGA_S		72.30	58.76		75.42	57.79		72.16	61.25		83.86	62.55
avg seq_ID		23.16	6.25		27.31	6.24		22.39	6.87		28.51	7.76
avg R-Rs/template		54	9		51	35		37	31		29	24
# PDB/SCOP_IDs	397	195	9	572	160	160	837	191	289	671	117	197
# SCOP categories		3	7		4	29		4	58		4	86
AA : (below)												
A				1	1		12	2	6	75		34
C	6		2	8		3	21		13	1		1
D				1			6		1	14		11
E				3		1	10		7	12		2
F	111	78	3	67	42 ¹	5	103	72	3	7		1
G	44	32		44	32		55	32	2	49	32	1
H	7	5	1	14	5	8	27	5	14	23		5
I	1		1	19		9	44		21	16		9
K				2		2	3		3	23		12
L				11		5	29		15	45		14
M				2		2	1		1	2		2
N	224	80		224	80		222	79	1	244	81	18
P										10		10
Q				15		14	12		10	45		19
R				82		49	161		108 ²	35		23
S	1						2			4		1
T	1			15		4	43	1	16	13	4	4
V				54		50	70		59 ³	33		23
W							3		3	4		1
Y	2		2	10		8	13		5	16		6
X							1		1			
1	All F297 are HIV1.											
2	108 of R297 are described by scop as "nuclear receptor ligand-binding domain"											
3	59 of V297 are described by scop as " RbcX-like"											

Suppl. Table 4. Effect of window_size on sequence variability in SCOP-categorized structure fragments at position D328. For column headings see Suppl. Table 2.

Position: D328													
window size	window 90			window 80			window 70			window 50			
classification	PDB	e.8	not e.8	PDB	e.8	not e.8	PDB	e.8	not e.8	PDB	LGA > e.8 ¹	e.8	not e.8
avg LGA_S		79.28	66.86		81.67	58.64		83.96	63.32		100.00	84.42	74.18
avg seq_ID		29.26	21.79		28.86	7.08		30.73	8.39		100.00	34.78	24.32
avg R-Rs/template		60	53		53	34		42	32		50	35	27
# PDB/SCOP_ID	277	119	1	304	124	17	333	120	34	230	2	73	1
# categories		3	1		3	10		3	15		1	3	1
AA: (below)													
A							3		2				
C													
D	274	119	1	278	124	1	270	120	1	226	2	73	1
E													
F													
G	2			2			3			2			
H										1			
I				2		2	2		1				
K				4		4	3		3				
L				13		9	34		21 ²				
M													
N				1		1	1		1				
P							3						
Q													
R													
S													
T										1			
V	1			4			14		5 ³				
W													
Y													
X													
¹ This exact match to a non-e.8 SCOP_ID is "b.47.1.4, viral cysteine protease of trypsin fold". This is a component of an e.8 protein but was selected here because its corresponding e.8 protein (poliovirus) was not selected at window 50.													
² Biotin holoenzyme synthetase (12), Ohr/OsmC resistance proteins (7), Acetyl-CoA synthetase-like (2)													
³ DNAPolymerase beta, N-terminal domain (3), NTF2-like (2)													

Suppl. Table 5A. Effect of window_size on sequence variability in SCOP-categorized structure fragments at position L374. For column headings see Suppl. Table 2.

Position: 374												
window size	window 90			window 80			window 70			window 50		
classification	PDB	e.8	not e.8									
avg LGA_S		77.83	100.00		82.41	86.06		82.99	88.48		83.00	100.00
avg seq_ID		31.03	100.00		31.19	74.07		36.7	75.56		41.23	100.00
avg R-Rs/template		63	90		56	63		49	57		33	50
# PDB/SCOP_ID	212	77	2	216	80	3	216	83	3	144	59	2
# categories		1	1		1	2		2	2		1	1
AA: (below)												
A												
C	149	42		149	42		149	41		91	28	
D												
E												
F												
G												
H												
I												
K												
L	62	35	2	67	38	3	65	41	3	52	31	2
M								1				
N												
P												
Q												
R												
S												
T												
V												
W												
Y												
X	1						1			1		

Suppl. Table 5B. Position L374.

position, window	aver LGA	aver ID	# matches	# templates	scop ID	description	cys	leu	met
position 374, w50_5				N=61					
	83.00	41.23	33	59	e.8.1.4	RNA-dependent RNA-polymerase	28	31	
	100.00	100.00	50	2	b.47.1.4	Viral cysteine protease of trypsin fold		2	
position 374, w70_5				N=86					
	83.29	36.94	49	82	e.8.1.4	RNA-dependent RNA-polymerase	41	41	
	58.12	16.67	34	1	e.8.1.2	Reverse transcriptase			1
	100.00	100.00	70	2	b.47.1.4	Viral cysteine protease of trypsin fold		2	
	65.44	26.67	32	1	i.7.1.1	Reovirus components		1	
position 374, w80_5				N=83		count			
	82.41	31.19	56	80	e.8.1.4	RNA-dependent RNA-polymerase	42	38	
	100.00	100.00	80	2	b.47.1.4	Viral cysteine protease of trypsin fold		2	
	58.17	22.22	30	1	i.7.1.1	Reovirus components		1	
position 374, w90_5				N=79					
	77.83	31.03	63	77	e.8.1.4	RNA-dependent RNA-polymerase	42	35	
	100.00	100.00	90	2	b.47.1.4	Viral cysteine protease of trypsin fold		2	

Suppl. Table 7. Effect of window_size on sequence variability in SCOP-categorized structure fragments at position H413. For column headings see Suppl. Table 2.

Position: H413												
window size	window 90			window 80			window 70			window 50		
classification	PDB	e.8	not e.8									
avg LGA_S		90.20	55.16		85.88	56.63		88.89	59.58		86.81	66.72
avg seq_ID		54.39	5.36		45.83	8.18		48.91	7.44		43.20	8.55
avg R-Rs/template		78	36		63	35		56	32		38	26
# PDB/SCOP_ID	67	37	1	309	45	54	951	42	392	1316	50	695
# categories		1	1		1	21		2	89		2	148
AA: (below)												
A				9		1	146		73	120	1	59
C	1			11		2	25		8	7		
D				4		3	17		5	85		76
E				4		2	23		16	50		21
F				1			4		1	8		4
G	1			6			24		1	42		14
H	29	17		29	17		50	17	7	40	17	11
I	1			22		6	127		77	215		101
K	10	10		12	10	2	42	10	7	100	23	39
L				80		14	133		42	172		98
M				40			75		8	13		3
N							36		30	33		19
P										7		7
Q	20	9		17	9		25	9		86	9	28
R	2			5	1		21		8	40		30
S	1			16	3	2	67	1	24	120		83
T							21		14	28		13
V	1	1		49	5	18	94	3	60	86		42
W							9		7	47		37
Y	1		1	4		4	11	2	4	16		10
X							1		1	1		