SUPPLEMENTAL MATERIAL

ARCHITECTURE OF A FULL-LENGTH RETROVIRAL INTEGRASE MONOMER AND DIMER, REVEALED BY SMALL ANGLE X-RAY SCATTERING AND CHEMICAL CROSS-LINKING

Ravi S. Bojja[1]*, Mark D. Andrake[1]*, Steven Weigand[2], George Merkel[1], Olya Yarychkivska[1], Adam Henderson[1], Marissa Kummerling[1] and Anna Marie Skalka[1]**

[1]Institute for Cancer Research, Fox Chase Cancer Center, 333 Cottman Ave., Philadelphia, PA 19111 and [2]DND-CAT Synchrotron Research Center, Northwestern University, APS/ANL 432-A004, 9700 South Cass Avenue, Argonne, IL 60439

Running head: Apo-integrase solution architecture

** Address correspondence to: Anna Marie Skalka, Institute for Cancer Research, Fox Chase Cancer Center, 333 Cottman Avenue, Philadelphia, PA 19111. Tel: 215 728 2490; Fax: 215 728 2778; E-mail: AM_Skalka@fccc.edu

***These authors contributed equally to this work**

SUPPLEMENTAL METHODS

Protein Expression and Purification. Cloning, bacterial expression, and purification of ASV IN and its derivatives has been described in previous publications (1,2). For isotopic labeling, the IN gene was inserted into the NdeI/HindIII restriction sites of the p11 vector (Structural Genomics Consortium, University of Toronto), which contains an N-terminal His-tag with a TEV protease cleavage site (3). The resulting plasmid was expressed in BL21 DE3 cells that were grown in an optimized M9 medium supplemented with all unlabeled amino acids except lysine and arginine, which were replaced with 1 mM of L-Lysine (U-13C6, 97-99%; U-15N2, 97-99%) and 1 mM of L-Arginine (U-13C6, 97-99%; U-15N4) (Cambridge Isotope Laboratories, Inc). MS/MS analyses showed that the extent of incorporation of the isotopically labeled amino acids was 95% and 90% respectively (supplemental Fig. S2). The proteins were purified as described below except that before the heparin column step, the His-tag was removed with TEV protease (3).

Standard protocol for ASV IN purification. Proteins were produced and purified as described in the following typical example: A 1-liter culture of *E. coli* BL21(DE3) cells containing the expression plasmid is grown to an optical density of 1.0–1.2 at 600 nm. Cells are then induced by addition of IPTG to 1 mM and harvested 3 hr postinduction by centrifugation at 10,000g for 10 min 4 °C. Cells are then suspended to a concentration of 6 ml/g wet cell paste in lysis buffer (50 mM BisTris pH 6.5, 1 M NaCl, 1 M Urea, 5 mM Immidazole, 5% glycerol, 6 mM 2-mercaptoethanol, and protease inhibitors; aprotinin, leupeptin, pepstatin, and phenylmethanesulfonyl fluoride (PMSF from Sigma), and lysed by two passes through a French pressure cell at 18,000 psi. The lysate is subjected to centrifugation at 15,000 x g for 30 min, and the supernatant filtered (0.45 micron) prior to loading on an iminodiacetic acid (IDA)–Sepharose (HiTrap IDA) column charged with 50 mM $NiSO_4$ and equilibrated with lysis buffer. The column is washed with 5 column volumes of binding lysis buffer and the protein eluted with a gradient from 5 mM to 750 mM imidazole. The eluted fractions are collected into 0.4 mM EDTA (final concentration) to prevent metal-induced aggregation of the protein. The salt concentration of the IDA-purified protein fractions is adjusted to 200 mM and they are applied immediately to a heparin–Sepharose column (HiTrap heparin) equilibrated in binding buffer (50 mM BisTris pH 6.5, 0.2 M NaCl, 0.1 mM EDTA, 10% glycerol, 1% Thiodiglycol, 6 mM 2-mercaptoethanol). The column is washed with 5 column volumes of binding buffer followed by a 10 column volume exponential gradient of NaCl (0.25–1.2 M) and fractions containing pure IN are pooled, concentrated on Amicon filters (YM10) and subsequently dialyzed in 25 mM BisTris pH6.1, 500 mM NaCl, 0.1 mM TCEP, 0.1 mM EDTA, 5% glycerol and stored at -70 °C. As an alternate to dialysis, some preparations include a final step of size exclusion chromatography on a Superdex 200 column, followed by concentration and flash freezing in liquid nitrogen.

Assays for IN catalytic activities. Assays for measuring the processing and joining activities have been described in detail previously (2,4). Concerted integration was assayed according to methods established for HIV IN (5,6), with the following modifications: final reaction conditions in a 25 or 50 microliter volume were 20 mM Hepes pH 7.5, 5 mM DDT, 10% PEG 3.35K, 20 μM $ZnSO_4$, 30 mM $MgCl_2$, 10 nM DNA donor, 10 μg/ml ΦX 174 RF I target DNA, and 80 nM ASV IN, for 1 to 2 hrs at 37 °C. The reaction was stopped by adding EDTA to a final concentration of 50 mM and SDS to 0.5%, then treating with Protease K at 400ug/ml final concentration for 60 min at 37 °C. Aliquots of each reaction were run on a 0.8% Agarose gel, with a 1X TBE/1M Urea buffer at 80V for 2 hrs and stained with Syber Green.

Fitting the atomic resolution data and cross-linking results with the SAXS determined dimer envelope. Docking of the ASV integrase dimer into the SAXS determined envelope was performed by using the data-driven biomolecular docking software HADDOCK v2.0 (7). The starting monomer IN configuration for the docking was constructed from the two domain structure of ASV IN (PDB code: 1C1A) with addition of the ASV IN NTD modeled from the coordinates of the HIV 1-212 (PDB code: 1K6Y) using a fully automated protein structure homology-modeling server at SWISS-MODEL (8,9). HADDOCK was performed on the monomer structures taking all residues into consideration as well as distance constraints imposed by our chemical cross-linking data. All lysine's observed in the cross-linking were defined in the ambiguous interaction constraints (AIRs) distance tbl file with a minimum of 2.5 Å distance to a maximum of 11 Å distance between the observed hybrid adducts. The initial run was performed with rigid CTD linkers and flexible NTD linkers in the docking monomers at default parameters in expert interface. The resulting minimum structure was further refined by a final run at the Guru interface with imposed C2 symmetry on each docking monomer with the following docking parameters: Residues 1-41, 60-199, and 224-268 of the NTD, CCD, and CTD, respectively were defined as semi-flexible regions of the docking partners, while residues 42-58 and 200-223 were allowed as fully flexible motifs.

During the rigid-body energy minimization, 1,000 structures were calculated with an option of cross-docking between all the randomly generated docking structures based on distance constraints. For each of the 1,000 combinations, 3 rigid-body docking trials were performed, and structures with minimum energy were further refined into 200 energy minima structures. The 200 best solutions based on the intermolecular energy were used for semiflexible simulated annealing, followed by a refinement in explicit water. Finally, the solutions were clustered by using default 7.5 Å rmsd based on the pairwise backbone rmsd matrix to the starting monomer.

SUPPLEMENTAL REFERENCES

1. Andrake, M. D., Ramcharan, J., Merkel, G., Zhao, X. Z., Burke, T. R., Jr., and Skalka, A. M. (2009) *AIDS Res Ther* **6**, 14
2. Merkel, G., Andrake, M. D., Ramcharan, J., and Skalka, A. M. (2009) *Methods* **47**, 243-248
3. Tropea, J. E., Cherry, S., and Waugh, D. S. (2009) *Methods Mol Biol* **498**, 297-307
4. Chow, S. A. (1997) *Methods* **12**, 306-317
5. Li, J., Dai, Z., Jana, D., Callaway, D. J., and Bu, Z. (2005) *J Biol Chem* **280**, 37634-37643
6. Li, M., Mizuuchi, M., Burke, T. R., Jr., and Craigie, R. (2006) *Embo J* **25**, 1295-1304
7. de Vries, S. J., van Dijk, M., and Bonvin, A. M. (2010) *Nat Protoc* **5**, 883-897
8. Arnold, K., Bordoli, L., Kopp, J., and Schwede, T. (2006) *Bioinformatics* **22**, 195-201
9. Kiefer, F., Arnold, K., Kunzli, M., Bordoli, L., and Schwede, T. (2009) *Nucleic Acids Res* **37**(Database issue), D387-D392

SUPPLEMENTAL FIGURE LEGENDS

Fig. S1. *Ab initio* shape models for dimer and monomer ASV IN. To test for uniqueness, 10 shape reconstructions were performed with the program GASBOR, for dimers, *A*, and monomers, *B,* using input files generated with an increasing qmax range of between 0.4 to 0.9 A^{-1}. These shapes were then subjected to averaging using the program DAMAVER, and the resulting normalized spatial discrepancy (NSD) values are shown below each shape. In each panel, the averaged filtered shape is shown in the top left, and the atomic resolution model fit within the envelope is shown on the bottom left. The dimer model shape in the bottom right in *A* (NSD=1.08) was chosen for refining cluster models during the HADDOCK process, as it bore the closest resemblance to the most frequently found clusters, and it is the shape portrayed in Figures 3 and 7.

Fig. S2. Isotopically labeled ASV wild type integrase (IN). GHM is an addition of three amino acids from the expression plasmid (p11). Stably incorporated isotopes of lysine (K+8) and arginine (R+10) residues are highlighted in blue and red respectively. The sequence coverage is more than 95%, with 19 of the 20 lysines and 18 of the 20 arginine residues observed with high confidence.

Fig. S3. Representative mass spectrometry data. *A*. CTD-CTD dimer linkages between labeled and unlabeled IN monomers. The insert shows a hybrid adduct corresponding to cross-linked peptide precursor ion (MW=2678.6Da) derived from the unlabeled IN sequence KVK(blue) and labeled IN sequence JVJPDITQJDEVTJJ (green). Labeled lysines are shown as "J" and lysine residues involved in cross-linking are shown by ($1). *B*. Mass spectrometry data for NTD-core dimer linkages show an extensive ion series for the hybrid precursor ion (MW=3328.6 Da) derived from the cross-linked peptides, with sequence U($1)HMPLR (blue) and AIJ($1)TDNGSCFTSJSTOEWLAO (green). Cross-linked residues are shown as ($1), N-terminal glycine is shown as U, and further labeled lysine and arginine are shown as J and O respectively.

Fig. S4**.** Iterative HADDOCK docking**.** *A*. Statistical analysis of HADDOCK results for the three iterative runs minimizing the R_g for IN dimer formation. *B***.** The first iterative run was preformed with rigid CTD linkers and fully flexible NTD linkers, while the NTD, CTD, and core domains were allowed to be semi-flexible (i.e. secondary structural elements are preserved). From the first run, one of the minimum arrangements is shown as Step 1, with a unique interface mediated by tryptophan residues (W259) edge-to-edge orientation (space fill, magenta) with an estimated R_g of 47. Further iterative refinement of the architecture in Step 1, with constraints between interacting NTD and CTD at the interface, yielded the arrangement in Step 2, showing stacked tryptophans at the interface with a R_g of 41. Finally, C2 symmetry and tighter distance constraints were imposed on the Step 2 model to achieve that shown in Step 3, with a R_g of 37. All the domains are color coded as in Figure 1, and the direction of the colored arrow represents the iterative movement of the specific domain with the same color. As indicated by the arrow at the right, the number of matches among the 39 linked peptides that are accommodated at the fist and final docking steps increased from 22 to 34. The final step 3 cluster was the best fit model within the SAXS envelope.

Fig. S5. Potential stabilizing interactions in ASV and HIV-1 IN reaching dimers *A*. ASV IN. Hydrophobic (Hy) and aromatic (A) interactions between proximal W259 residues from each monomer are a primary stabilizing force in this reaching dimer interface. In addition, loop residues 244-246(RGY) from each monomer form direct hydrogen bonds (H) with the other monomer. The same RGY loop residues are also involved in hydrogen bonding with the NTD of the second monomer. Backbone bonds (bb) are shown as solid black lines and side chain hydrogen bonds as red dashed lines. *B*. HIV IN. This interface can be stabilized by H-bonding of the NH group of W243 to the backbone carbonyl of K244, as well as hydrophobic interactions between W243 side chains from each monomer. In addition, the R262 side chain from one monomer, can form hydrogen bonds with P30 and V31 of the second monomer.

Finally, the β-strand containing residues 257-259 from each monomer can come together in anti-parallel fashion to form a β-barrel structure comprising 3 strands from each CTD.

SUPPLEMENTAL TABLES

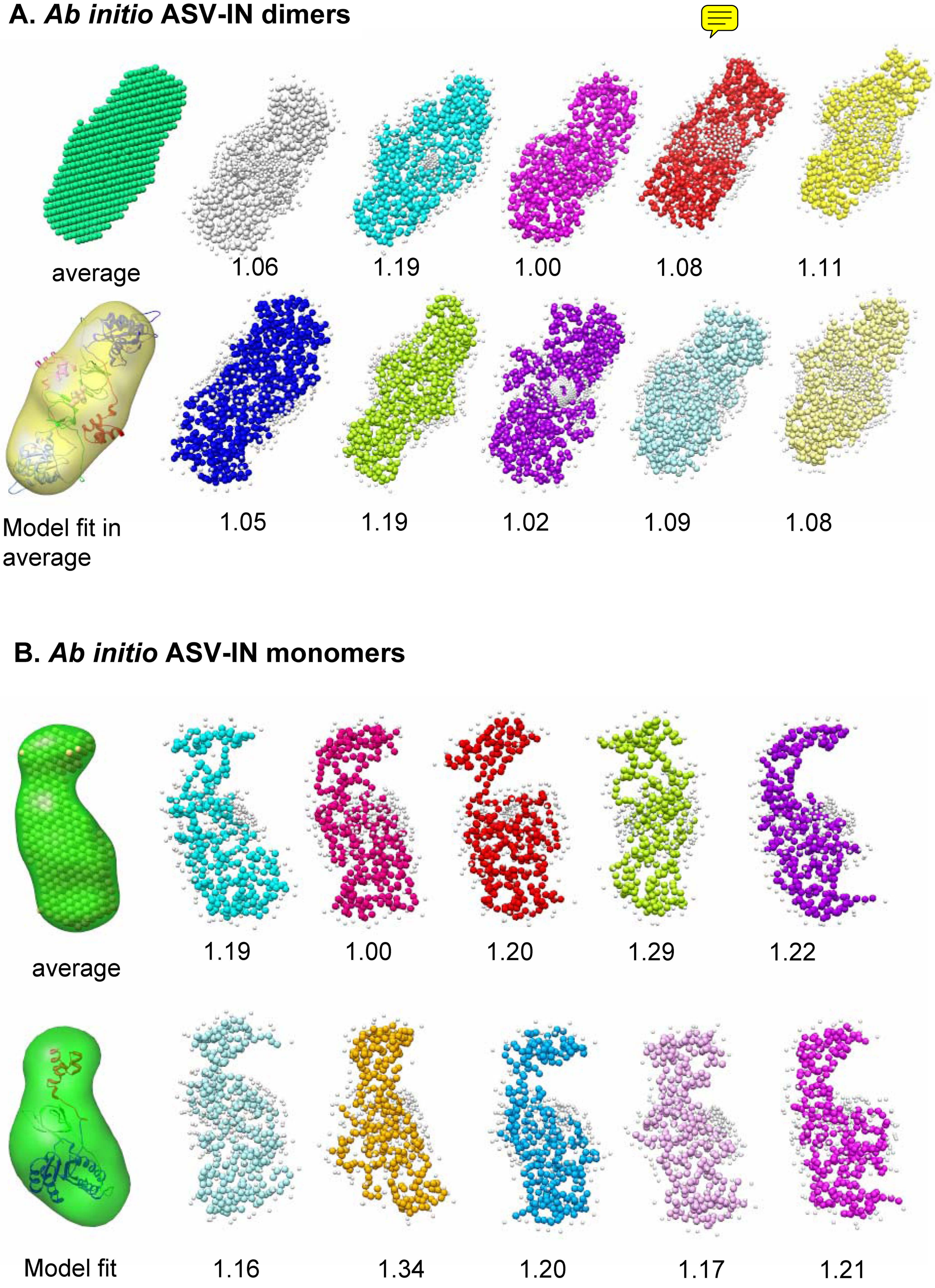
Table S1 Observed BS3 modified lysines in the ASV integrase (IN) in monomer band, cross-linked residues are shown as ($1).

Table S2 Observed BS3 cross-links in ASV integrase (IN) dimer band between labeled and unlabeled proteins, cross-linked residues are shown as ($1), labeled lysine is shown as J and labeled arginine is shown as O, N-terminus glycine involved in cross-linking is shown as U.

Table S3 Observed BS3 cross-links in the core region of the ASV integrase (IN) tetramer band between labeled and unlabeled proteins, cross-linked residues are shown as ($1), labeled lysine is shown as J and labeled arginine is shown as O, N-terminus glycine involved in cross-linking is shown as U.

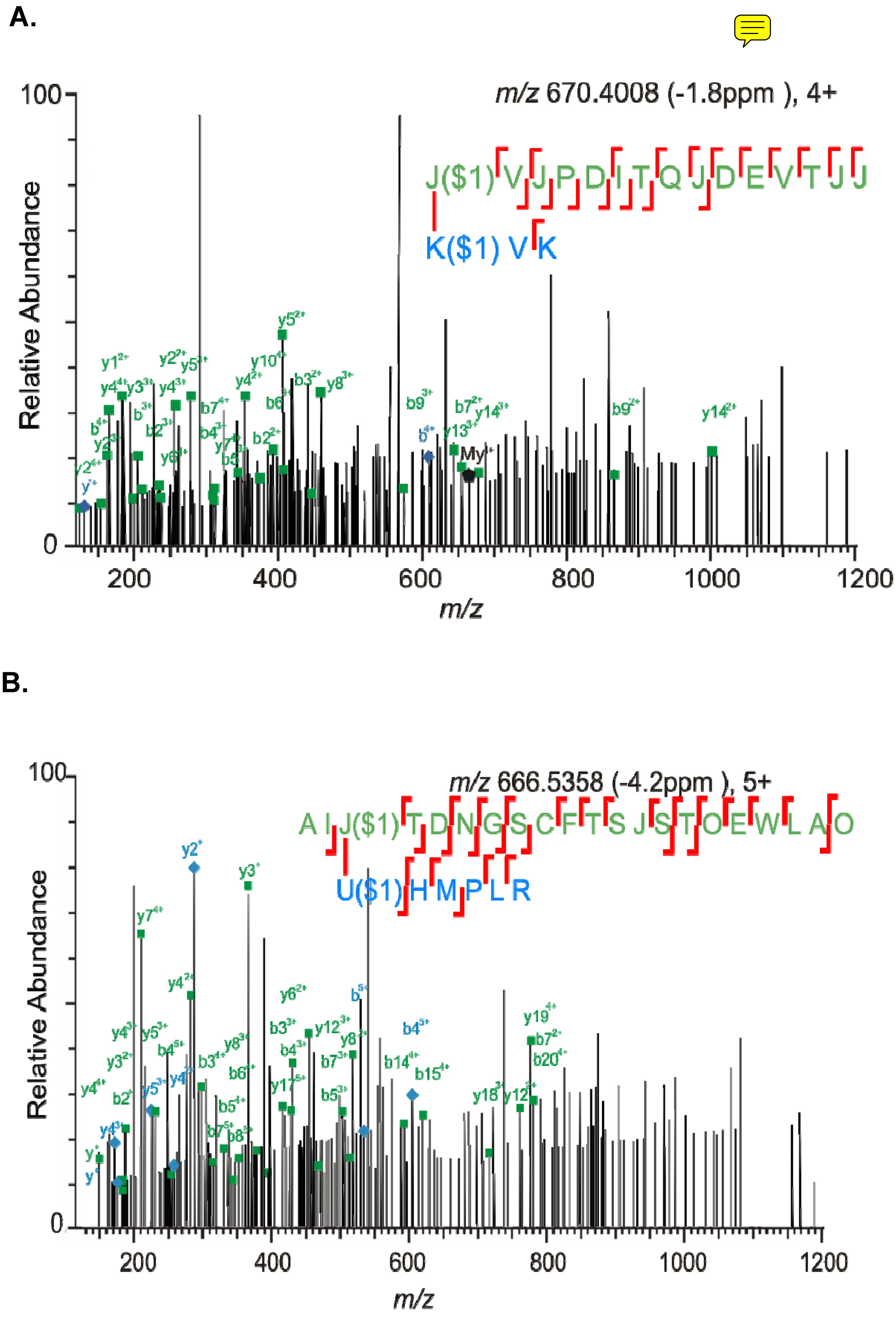
SUPPLEMENTAL Movies

1 – ASV IN reaching dimer model fit within SAXS envelope.
2 – Change in CTD position required to bind DNA.

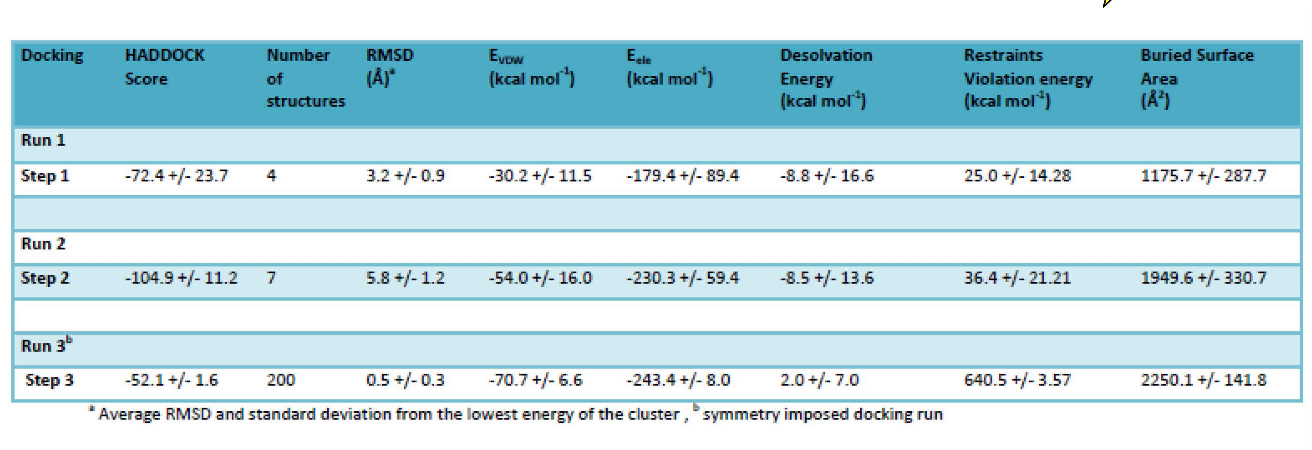

A. Ab initio ASV-IN dimers
average
1.06
1.19
1.00
1.08
1.11
Model fit in average
1.05
1.19
1.02
1.09
1.08
B. Ab initio ASV-IN monomers
average
1.19
1.00
1.20
1.29
1.22
Model fit
1.16
1.34
1.20
1.17
1.21

ISOTOPICALLY LABELED WILD TYPE ASV INTEGRASE

```
             10         20         30         40         50
GHM  PLREAKDLHT ALHIGPRALS KACNISMQQA REVVQTCPHC NSAPALEAGV

             60         70         80         90        100
     NPRGLGPLQI WQTDFTLEPR MAPRSWLAVT VDTASSAIVV TQHGRVTSVA

            110        120        130        140        150
     AQHHWATAIA VLGRPKAIKT DNGSCFTSKS TREWLARWGI AHTTGIPGNS

            160        170        180        190        200
     QGQAMVERAN RLLKDKIRVL AEGDGFMKRI PTSKQGELLA KAMYALNHFE

            210        220        230        240        250
     RGENTKTPIQ KHWRPTVLTE GPPVKIRIET GEWEKGWNVL VWGRGYAAVK

            260        270        280
     NRDTDEVIWV PSRKVKPDIT QKDEVTKKDE ASPLFA
```

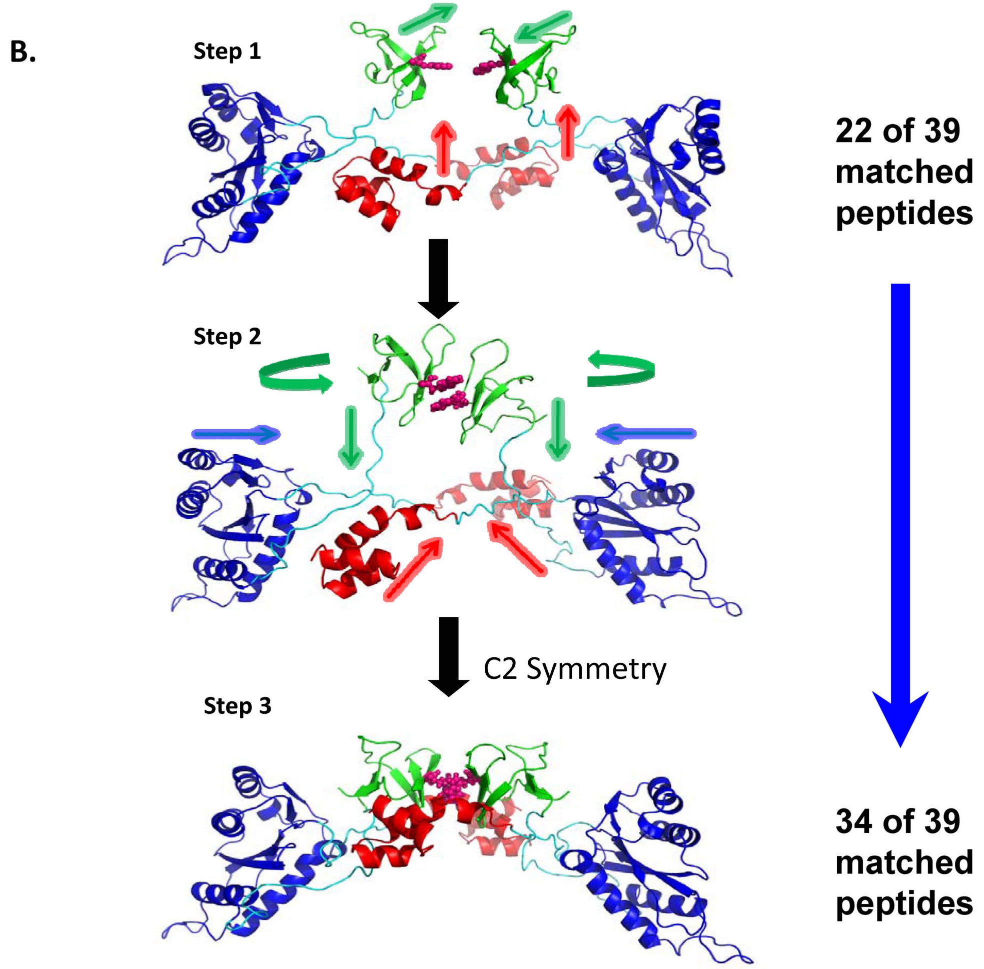

A.
m/z 670.4008 (-1.8ppm), 4+
J($1) V J P D I T Q J D E V T J J
K($1) V K
Relative Abundance
100
0
200
400
600
800
1000
1200
m/z
y1²⁺
y4⁴⁺
y3³⁺
y2²⁺
y5³⁺
y4³⁺
y10⁴⁺
y5²⁺
y4²⁺
b3²⁺
y8³⁺
b⁴⁺
b³⁺
b7⁴⁺
b6³⁺
y2³⁺
b2³⁺
b4³⁺
b2²⁺
y7⁴⁺
b5³⁺
y6⁴⁺
y2⁴⁺
y⁺
b9³⁺
b7²⁺
y14³⁺
y13³⁺
b⁴⁺
My⁴⁺
b9²⁺
y14²⁺
B.
m/z 666.5358 (-4.2ppm), 5+
A I J($1) T D N G S C F T S J S T O E W L A O
U($1) H M P L R
Relative Abundance
100
0
200
400
600
800
1000
1200
m/z
y2⁺
y3⁺
y7⁴⁺
y4²⁺
y6²⁺
b⁵⁺
y4³⁺
y5³⁺
b4⁵⁺
y3²⁺
b3⁴⁺
y8³⁺
b3³⁺
y12³⁺
b4³⁺
y8²⁺
b7³⁺
b4³⁺
b14⁴⁺
b15⁴⁺
y19⁴⁺
b7²⁺
b20⁴⁺
y4⁴⁺
b2⁺
y5³⁺
y4²⁺
b6⁴⁺
b5⁴⁺
y17⁵⁺
b5³⁺
y18³⁺
y12²⁺
y4³⁺
b7⁵⁺
b8⁵⁺
y⁺
y⁺

A.

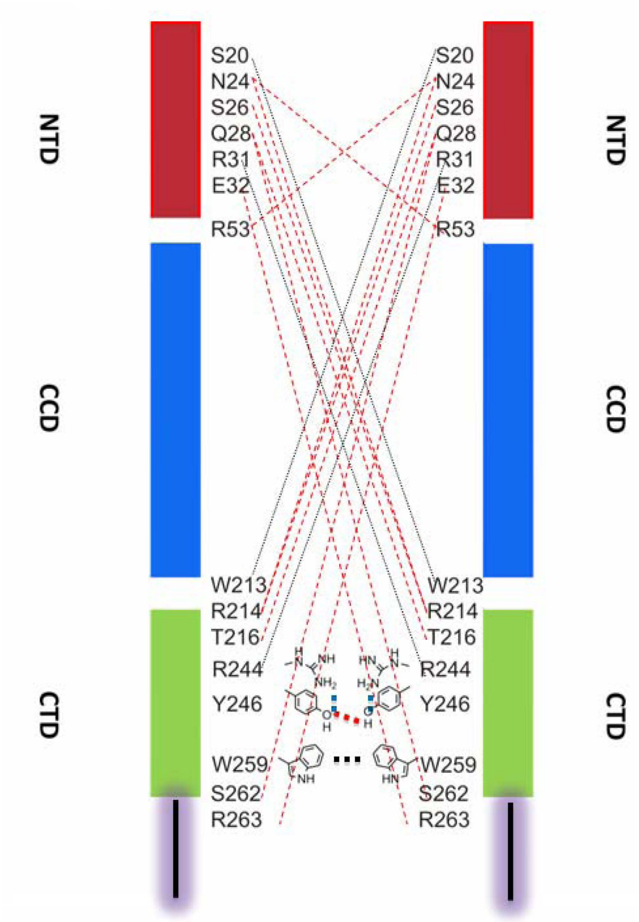


Docking	HADDOCK Score	Number of structures	RMSD (Å)[a]	E_{VDW} (kcal mol^{-1})	E_{ele} (kcal mol^{-1})	Desolvation Energy (kcal mol^{-1})	Restraints Violation energy (kcal mol^{-1})	Buried Surface Area (Å^2)
Run 1								
Step 1	-72.4 +/- 23.7	4	3.2 +/- 0.9	-30.2 +/- 11.5	-179.4 +/- 89.4	-8.8 +/- 16.6	25.0 +/- 14.28	1175.7 +/- 287.7
Run 2								
Step 2	-104.9 +/- 11.2	7	5.8 +/- 1.2	-54.0 +/- 16.0	-230.3 +/- 59.4	-8.5 +/- 13.6	36.4 +/- 21.21	1949.6 +/- 330.7
Run 3[b]								
Step 3	-52.1 +/- 1.6	200	0.5 +/- 0.3	-70.7 +/- 6.6	-243.4 +/- 8.0	2.0 +/- 7.0	640.5 +/- 3.57	2250.1 +/- 141.8

[a] Average RMSD and standard deviation from the lowest energy of the cluster , [b] symmetry imposed docking run

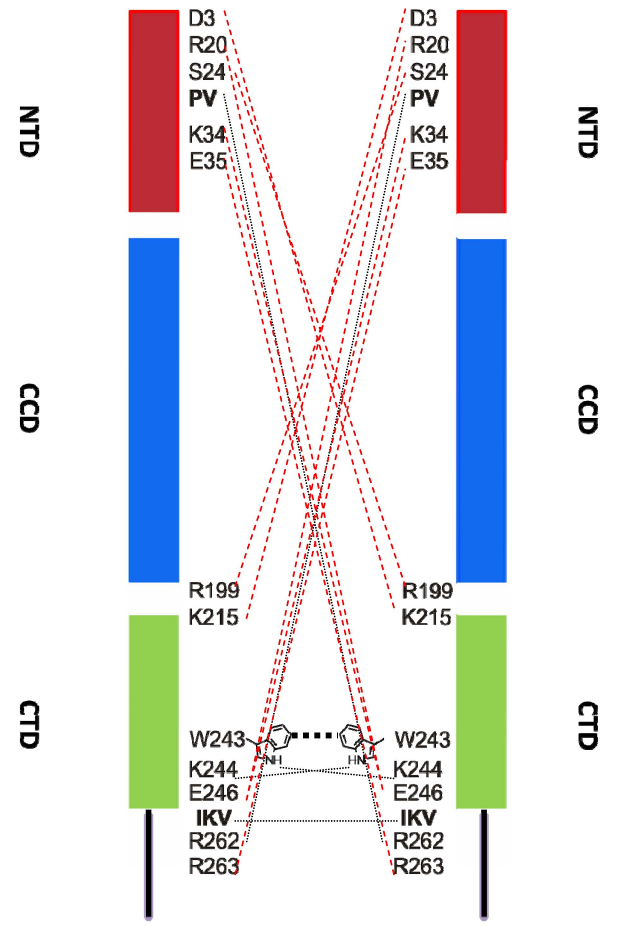


A.



ASV	ASV	type
W259	W259	Hy, A
Y246	Y246	H
R244	Y246	H
S20	W213	bb
R31	R244	bb
N24	R53	H
R214	N24,S26	H
Q28	T216,S262	H
E32	R263	H

B.



HIV	HIV	type
W243	W243	Hy,A
W243	K244	bb
K34	E246	bb,H
K215	D3	H
K258	I257,K258, V260	bb
K258	D256	H
R199	S24	H
R20	E246	H
R263	E35	H
R262	P30,V31	bb