Probability Analysis of Variational Crystallization and Its Application to gp120, The Exterior Envelope Glycoprotein of Type 1 Human Immunodeficiency Virus (HIV-1)*

(Received for publication, June 18, 1998, and in revised form, October 29, 1998)

Peter D. Kwong‡, Richard Wyatt§, Elizabeth Desjardins§, James Robinson¶, Jeffrey S. Culp¶, Brian D. Hellmig, Raymond W. Sweet, Joseph Sodroski§, and Wayne A. Hendrickson**

From the ‡Department of Biochemistry and Molecular Biophysics and **Howard Hughes Medical Institute, Columbia University, New York, New York 10032, the §Division of Human Retrovirology, Dana-Farber Cancer Institute, Boston, Massachusetts 02115, the ¶Department of Pediatrics, Tulane University School of Medicine, New Orleans, Louisiana 70112, and §SmithKline Beecham Pharmaceuticals, King of Prussia, Pennsylvania 19406

The extensive glycosylation and conformational mobility of gp120, the envelope glycoprotein of type 1 human immunodeficiency virus (HIV-1), pose formidable barriers for crystallization. To surmount these difficulties, we used probability analysis to determine the most effective crystallization approach and derive equations which show that a strategy, which we term variational crystallization, substantially enhances the overall probability of crystallization for gp120. Variational crystallization focuses on protein modification as opposed to crystallization screening. Multiple variants of gp120 were analyzed with an iterative cycle involving a limited set of crystallization conditions and biochemical feedback on protease sensitivity, glycosylation status, and monoclonal antibody binding. Sources of likely conformational heterogeneity such as N-linked carbohydrates, flexible or mobile N and C termini, and variable internal loops were reduced or eliminated, and ligands such as CD4 and antigen-binding fragments (Fabs) of monoclonal antibodies were used to restrict conformational mobility as well as to alter the crystallization surface. Through successive cycles of manipulation involving 18 different variants, we succeeded in growing six different types of gp120 crystals. One of these, a ternary complex composed of gp120, its receptor CD4, and the Fab of the human neutralizing monoclonal antibody 17b, diffracts to a minimum Bragg spacing of at least 2.2 Å and is suitable for structural analysis.

In conventional crystallizations of biological macromolecules, the protein or other macromolecular subject is treated as a fixed entity to be tested in a multitude of crystallization conditions. Despite advances such as sophisticated screening procedures (1, 2) and crystallization robots (3, 4), this approach often fails for components from complex biological systems. One of these, the subject of this study, is the HIV-1* exterior envelope glycoprotein, gp120. In such cases, success may follow if the protein itself is varied. There are, however, many options in this vein and it is not clear how they might be prioritized. By way of background for this study, we first consider various options for the crystallization of conformationally complex macromolecules and then describe the characteristics of gp120.

For the more difficult crystallization challenges, which can be defined as those for which conventional screening fails, one typically tries to vary or modify the protein while maintaining biologically important properties. Meaningful results are obtained since the integrity of internal structure and functional properties can often tolerate variation at the molecular surface where lattice contacts are made. The probability for success in crystallization is enhanced because flexible or heterogeneous surface features may be removed or because of the fortuitous introduction of lattice interactions. A prescient example that pre-dates the powerful methods of modern molecular biology was Kendrew’s (5) screening of myoglobins from many different organisms until he found one, from sperm whale, that crystallized well. Indeed, human myoglobin requires a Lys to Arg substitution in order to produce crystals suitable for structural analysis (6). Conversely, crambin forms exceptionally well ordered crystals despite being a mixture of two isoforms with sequence variation at internal residues (7).

There are many notable examples of variation or modification in the crystallization of macromolecules. Systematic variation in the species of origin, as pioneered with myoglobin (5), was also instrumental in the crystallization of the transcription initiating TATA-binding protein (8). Proteolysis is often used to define crystallizable fragments, following the early examples from enzymatic digestions of antibodies that produced crystallizable fragments (reviewed in Ref. 9) and the bromelain release of hemagglutinin from the influenza virus membrane (10). Variation of recombinant constructs, often inspired by proteolytic definition, is now commonplace with the widespread use of molecular biology tools. Systematic variation in the length of DNA oligomers has proved essential in the structural studies of protein-nucleic acid complexes. The work of Jordan and Pabo (11) on λ repressor sets the example for transcription factors, and the principle extends to other complexes such as the nucleosome (12). The use of protein ligands to stabilize another protein of interest for crystallization has also been effective as in the study of actin through its complex with DNase I (13) and more generally through complexes with antigen-binding Fab fragments of antibodies (reviewed in Ref. 14). The principle that the detergent-solubilized lipid interface of membrane proteins is generally unavailable for lattice contacts has led to the concept that crystallizability will be enhanced if the non-variable surface area is increased, and this was demonstrated in practice in the crystallization of a bacterial cytochrome oxidase in complex with an antibody Fv.

* This work was supported by grants from the Aaron Diamond Foundation, the National Institutes of Health, and the G. Harold and Leila Mathers Foundations, the Friends 10, the late William F. McCarty-Cooper, and Douglas and Judith Krupp. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

© 1999 by The American Society for Biochemistry and Molecular Biology, Inc.

© 1999 by The American Society for Biochemistry and Molecular Biology, Inc.

http://www.jbc.org/
fragment (15). Similarly, the anticipated conformational and compositional heterogeneity in carbohydrate moieties of glycoconjugates is expected to interfere with crystallization, and deglycosylation has proved essential for heavily glycosylated proteins such as human chorionic gonadotropin (16).

HIV-1 induces acquired immunodeficiency syndrome in humans (17, 18). The gp120 glycoprotein helps to mediate virus entry into cells through sequential recognition of two cellular receptors, the surface glycoprotein CD4 (19, 20) and a chemokine receptor (primarily CXCR4 or CCR5, depending on viral strain) (21–26). These high affinity interactions are attractive targets for mimetic drug design. Although the structure of the gp120-binding domain of CD4 and the identity of residues critical to its interaction with gp120 have been known for several years (27, 28), this has not been sufficient for design of potent antagonists (29–31). As the major virus-specific antigen accessible to neutralizing antibodies, knowledge of the gp120 structure could also impact considerably on vaccine design. Despite this interest and considerable effort for several years with pure soluble protein, available in quantities as a by-product in part from vaccine trials, gp120 has resisted crystallographic analysis. The mature gp120 glycoproteins of different HIV-1 strains typically have 470–490 amino acid residues (32). Extensive N-linked glycosylation at 20–25 sites accounts for roughly half of the gp120 mass (32, 33). Sequences from many different viral isolates show that gp120 has five variable regions (V1-V5) (32, 34) and nine conserved disulfide bridges (33). Except for limited N- and C-terminal cleavage, proteolytic digestion does not reveal a subdomain structure. Indeed, even after extensive proteolytic cleavage, the unreduced protein runs near its native mobility on SDS-PAGE. 

The gp120 glycoprotein likely exhibits conformational flexibility. Some of the variable regions, the V2 and V3 loops in particular, are known to be exposed on the surface of the native protein and probably assume multiple conformations. The potential of gp120 to undergo conformational change is also evidenced by shedding, the CD4-induced dissociation of gp120 from the surface of the virus, by ligand-induced variations in monoclonal antibody binding (35, 36), and by complex CD4-gp120 binding kinetics (37). These changes may be related to the functional role of gp120 in virus entry.

The extensive glycosylation and conformational heterogeneity of gp120 suggested that merely screening the protein through ever more exotic crystallization conditions would not produce well diffracting crystals. We have analyzed the effectiveness of optimizing different crystallization factors given the specific characteristics of gp120. This led us to a strategy employing radical modification of the protein surface, primarily to reduce heterogeneity but also to create new potential lattice contacts. We derive equations which show that this strategy, which we term variational crystallization, substantially enhances the overall probability of crystallization for gp120. An iterative process, involving both biochemical and molecular biological techniques, was used to detect and remove chemical and conformational heterogeneity. In addition, protein ligands, namely CD4 and the Fab fragments of several monoclonal antibodies, were used to restrict conformational mobility. Progressive trials of 18 different gp120 crystallization variants yielded six different crystals, at least one of which is suitable for structural analysis. This paradigm of crystallization, with a focus on protein modification rather than on crystallization screening, may aid in the structural analysis of other conformationally complex proteins.

**Theory**

Much of the crystallization literature is anecdotal, reflective perhaps of the diverse nature of proteins. Systematic quantitative studies have necessarily focused on robust, well characterized systems (38). If a particular protein fails to crystallize, one is faced with a bewildering array of options based on experience with other often quite different proteins. In the absence of a comprehensive crystallization theory it is difficult to know how to proceed. Here, we devise an approximate theoretical underpinning for such decisions based on the ratio comparing crystallization probabilities before $(p)$ and after $(p_{\text{enh}})$ a modifying procedure. We define the enhancement in crystallization probability as $\epsilon = p_{\text{enh}}/p - 1$, whereby $\epsilon = 0$ for no change and can reach a maximum, $\epsilon_{\text{max}} = 1/p_{\text{enh}} - 1$, that depends on the inverse of the initial probability.

In evaluating different crystallization strategies, one important consideration is effectiveness. Many factors affect crystallization, and a suitable crystallization approach depends on identifying and dealing with those that are most limiting. For example, if a protein were only 30% pure, the crystallization probability associated with such protein purity would be low and a purification strategy would be key; if a protein were 98% pure, further purification would most likely have little impact on the overall probability of crystallization. Factors that might be expected to affect the crystallization of gp120 are listed in Table I, along with estimates of the effect of optimizing each factor given the specific characteristics of gp120.

Although identification of limiting crystallization factors can establish rough guidelines as to the appropriateness of a particular crystallization strategy, a better way to evaluate effectiveness (or perhaps to judge the progress of a specific crystallization effort) is by quantitative assessment of the enhancement in crystallization probability. For example, if 80% of all crystallizable proteins crystallize from a core set of 50 conditions (2), a strategy that involves screening ever larger arrays of crystallization conditions could at most enhance the probability of crystallization by only 25% over that for the first 50 conditions; further screening would yield increasingly diminishing returns. With this screening example, the quantitative enhancement of probability is straightforward to calculate, but it is not immediately apparent for the strategy of variational crystallization, which focuses on protein modification. Here we consider two kinds of such modifications: those designed to reduce heterogeneity and those related to expanding the number of crystallization candidates.

Enhancement of Surface Homogeneity—Crystalline order is explicitly dependent on lattice homogeneity. Reducing heterogeneity can be thought of as increasing the proportion of surface area available for formation of lattice contacts, increasing the probability of crystallization. The probability that a single lattice contact between two molecules may form is in part related to the fraction of surface area that is homogeneous on one molecule multiplied by the fraction homogeneous on the other,

$$p(\text{homogeneous contact}) \times H(\text{molecule 1}) \times H(\text{molecule 2}) \quad (\text{Eq. 1})$$

where $H$ is defined as the homogeneous fraction of the surface. Consider the case where the molecule in question is the smallest repeating unit in a crystal, that is, the asymmetric unit. In such a case molecule 1 and molecule 2 are equal, and the above equation reduces to (% homogeneous surface)$^2$. Now consider the same scenario with two lattice contacts: the probability that both are homogeneous is related to $[H(\text{molecule 1})]^2 \times (H(\text{molecule 2}))^2$, where “$H$” is the homogeneous fraction of the surface which may form lattice contacts and “$\theta_n$” is a function of the relative size and total number of lattice contacts other than contact $n$.

---

2 P. D. Kwong, unpublished data.
Factors affecting the crystallization of gp120

| Protein characteristics | Specific gp120 characteristics | Optimization* |
|-------------------------|-------------------------------|---------------|
| A. Chemical homogeneity  | Very clean, affinity purification | +             |
| Other proteins          | Relatively limited, source dependent | ++            |
| Carbohydrate variation  | N- and C-terminal ragged | ++            |
| Polypeptide variation   | N-Linked, large and flexible | ++            |
| B. Conformational heterogeneity | Total of ~130 amino acids are flexible | +++           |
| Carbohydrate            | Roughly 30 amino acids are flexible | ++            |
| N or C termini          | Function-related conformational mobility | +++          |
| Other                   |                               |               |

Screening characteristics

| Protein | Solubility | Stability | Availability | Screening variations | Protein homologues | Ligands |
|---------|------------|-----------|--------------|----------------------|-------------------|---------|
| Solubility | Requires ~300 mM NaCl for solubility | ++ |               |                      |                   |         |
| Stability  | Stable for over 1 week in cell culture | + |               |                      |                   |         |
| Availability | Approximately 2 mg/liter of cell culture | + |               |                      |                   |         |
| Screening variations | Many different gp120 isolates | +++ |               |                      |                   |         |
| Protein homologues | Many ligands (dozens of monoclonal antibodies available) | +++ |               |                      |                   |         |

* Estimate of the effect on the crystallization probability of a strategy which optimizes the particular factor. The number of (+) symbols denotes the size of the effect; + refers to almost no change in probability after optimization, whereas ++++ refers to a large change in probability. The scale used here is a qualitative estimate; for more quantitative results, see Table III. For chemical heterogeneity, optimization refers to the effect on crystallization of making the protein more chemically homogeneous. For conformational heterogeneity, optimization refers to the effect of removing or circumventing the particular source of heterogeneity.

and the degree and distribution of surface homogeneity, related to the occlusion of available surface area upon formation of each lattice contact as well as the spatial distribution of homogeneous surface over the molecular surface. Generalizing to the case of “C” lattice contacts, the probability associated with homogeneous lattice formation is related to,

\[ p(\text{lattice}) = \frac{\prod_{n=1}^{c} (H-a_1)^2}{(H-a_1)^2} = \frac{c}{(H-a_1)^2} \] (Eq. 2)

In the restricted case of one molecule per asymmetric unit, the observed average value of “C” (\( C_{\text{ave}} \)) is ~4.5 (39), with a minimum theoretical value for the most common space groups of 2 or 3 (39). Since \( C \) may be relatively small, lattice contacts may make up only a small proportion of a macromolecule surface, with considerable surface heterogeneity tolerated. Thus, for example, many proteins that pack into well ordered crystal lattices have disordered regions, with N and C termini as well as internal loops being unresolved.

Given a reduction in surface heterogeneity, what is the change in crystallization probability? Surface area is correlated with molecular mass (M) by the power law: surface area \( = 6.3 \times M^{0.73} \), which on average predicts surface area to within 4% for monomeric proteins (40). The fraction of homogeneous surface can thus be approximated as a ratio of molecular masses of the total and of the homogeneous portion of the protein,

\[ H \approx \frac{\text{M(_homogeneous)}}{\text{M(total)}}^{0.73} \] (Eq. 3)

From Equations 2 and 3, it is now possible to estimate the enhancement in probability for crystallization upon reduction of heterogeneity. With the simplifying approximation \( a_n \approx 0 \), the probability ratio of before \((p_B)\) and after \((p_A)\) becomes,

\[ \frac{p_A}{p_B} = \frac{\text{M(_homogeneous)}}{\text{M(total)}}^{1.46-C} \] (Eq. 4)

Equation 4 is still not very useful, however, since \( \text{M(homogeneous)} \) is unknown and molecule-specific. In reducing heterogeneity, however, it seems reasonable to assume that the removed portion, if it were a highly branched carbohydrate or a proteolytically exposed region, is completely heterogeneous. In such cases, \( \text{M(_homogeneous)} \approx \text{M(homogeneous)} \), whether or not all heterogeneity has been removed. Assuming that \( C \approx C_{\text{ave}} \), the enhancement \((\epsilon_s)\) in probability on removal of a heterogeneous portion becomes,

\[ \epsilon_s = p_A/p_B - 1 = \frac{\text{M(total)}}{\text{M(total)}}^{1.46-C_{\text{ave}}} - 1 \] (Eq. 5)

This last equation allows the change in crystallization probability upon heterogeneity removal to be quantified. For example, consider a situation where a recombinant DNA approach is used to produce a protein with an affinity tag of 10 amino acid residues. Is it important to remove these presumably flexible residues? From Equation 5, the answer depends on the protein size. For a 100-residue protein, removal of the tag would greatly enhance the crystallization probability by: \( \epsilon_s = p_A/p_B - 1 \approx 0.87 \), whereas for a 500-residue protein the enhancement would be minimal, \( \epsilon_s = 0.14 \).

Another variant of Equation 4 can be used to estimate the impact of adding a ligand of fixed structure to a molecule that contains heterogeneous portions. This expands the surface available for lattice contacts and effectively dilutes the heterogeneous component. It may be an approach of choice when the heterogeneity is essentially unremovable, such as at the lipid interface of detergent-solubilized membrane proteins. One faces the difficulty of estimating the extent of heterogeneity to use Equation 4, but this might be done by summing residual variable components or by topographical estimates for a membrane protein. (For example, for a sphere embedded symmetrically in a membrane of thickness h, \( 1 - H = \text{area(heterogeneous/area(total))} = h/[6Mv/(\pi N_o^2)]^{1/3} \), where M is molecular mass, \( v \) is partial specific volume, and \( N_o \) is Avogadro’s number. Thereby, \( 1 - H = 0.62 \) for \( h = 30 \text{ Å} \) and \( M = 50 \text{ kDa} \).) Then the enhancement \((\epsilon_s)\) in probability on addition of a fixed component becomes,

\[ \epsilon_s = \frac{\text{M(total)}}{\text{M(total)}}^{1.46-C_{\text{ave}}} - 1 \] (Eq. 6)

In the instance of a 50-kDa protein, half of which is heterogeneous, to which a 25-kDa Fv fragment is complexed, \( \epsilon_s = [(50/25) \times (75-25)/(50-25)]^{1.46-4.5} - 1 \). Thus if the overall crystallization probability of the protein was initially only 1 chance in 10,
assuming all other crystallization probability components remained unchanged, the crystallization probability of the Fv fragment complex would be roughly 1 chance in 2, a substantial enhancement.

The accuracy of the quantification is only as good as the approximations, and several of the approximations used here call for further scrutiny. The approximation of molecular mass for surface area was used for the initial protein prior to heterogeneity removal. This is probably an underestimate since the category of heterogeneity is quite broad, and there are some situations, such as with segmental flexibility where these equations may be invalid. For example, in the case of two rigid domains connected by a flexible linker, one would have to consider the possibility that one domain could be fixed relative to the other with a single appropriate contact.

*Increase of Molecular Variants*—Another aspect of varia-
tional crystallization, the use of multiple variants of the same protein, also increases the probability of crystal formation. In this case, the overall probability of crystallization is exponentially related to the number of variants. Assuming independence of variants (a reasonable assumption with different protein ligands; not as valid with minor changes) with additional crystallization, the use of multiple variants of the same protein can benefit from a strategy employing multiple variants.

**EXPERIMENTAL PROCEDURES**

**Constructs of gp120**—The various recombinant gp120 glycoproteins used for crystallization trials were produced in stable *Drosophila* Schneider 2 lines under the control of an inducible promoter as described previously (41) (Table II). Genetic constructs containing various deletions and substitutions were made during the course of dissecting the gp120 domain structure. The procedures for making these constructs and the biological properties of the corresponding protein products are described elsewhere (see references in Table II).

**Protein Production and Purification**—The N-terminal two domains of CD4 (D1D2), residues 1–183, were produced in Chinese hamster ovary cells and purified as described previously (27). Human monoclonal antibodies 17b, A32, C11, and F105 (derived from HIV-infected individuals) (42, 43) and mouse monoclonal antibodies L71 and 178.1 (44, 45) were purified by protein-A affinity chromatography. Secreted gp120 from *Drosophila* cells was purified by affinity chromatography with the F105 antibody covalently coupled to Sepharose. Following extensive washing with phosphate-buffered saline containing 0.5 M NaCl, gp120 protein was eluted with 0.1 M glycine, pH 2.8, followed by immediate neutralization with Tris buffer.

**Protease Digestion**—Fab fragments were produced by papain digestion of monoclonal antibodies. Briefly, the antibody was reduced in 100 mM dithiothreitol, 100 mM NaCl, 50 mM Tris, pH 8.0, for 1 h at 37 °C, and dialyzed (4 °C), first in phosphate-buffered saline to reduce the dithiothreitol concentration to ~1 mM, then in alkylating solution (phosphate-buffered saline with 2 mM iodoacetamide, pH 7.5, 48 h), and subsequently in phosphate-buffered saline without iodoacetamide. The reduced and alkylated antibody was concentrated to at least 2 mg/ml and digested with papain using a commercial protocol (Pierce). An additional gel filtration chromatographic step on a Superdex S-200 column (Pharmacia, fast protein liquid chromatography) was added to ensure oligomeric homogeneity.

The gp120 proteins were subjected to digestion with papain, elastase, and subtilisin (Boehringer-Mannheim) to assay for proteolytic susceptibility. In these assays, the gp120 concentration was kept constant and the protease diluted serially (3.3 ×) from a ratio of 1.10 to 1.100. The digestion mixture was incubated for 1 h at 37 °C and quenched by adding 1% SDS (1:10 ratio) with immediate heating in boiling water for 2 min. Digestion products were analyzed with SDS-PAGE with and without dithiothreitol reduction.

Carboxypeptidase Y digestion was used to analyze the C terminus of gp120. A 1:10 ratio of carboxypeptidase Y (Boehringer-Mannheim) to gp120 was incubated for 1 h at 37 °C, pH 7.0. Even though digestion could not be easily seen by SDS-PAGE, the C terminus of gp120, HXBc2 strain, contains a number of positively charged amino acids, and the extent of the reaction could be monitored by native-PAGE.

**Deglycosylation**—*Drosophila* produced gp120 proteins were deglyco-
sylated enzymatically. Briefly, 0.5 mg/ml gp120 was incubated with various deglycosylating enzymes (singly or in combination) in 0.5 M NaCl, 100 mM sodium acetate, pH 5.7, for 10 h at 37 °C. Endoglycosi-
dase D was used at a concentration of 0.1 unit/ml, endoglycosidase F at

| Construct name | gp120 strain | Amino acid for construct | Ref. |
|---------------|-------------|-------------------------|-----|
| Δ31-III | IIIB | 62-511 | 60 |
| Δ30-FL | JRFL | 31-511 | 61 |
| ΔV1/2V3 | BH10/HXBc2 | 31-120 GAG 204-297 GAG 330-511 | 53 |
| ΔV1/2V3ΔC5 | BH10/HXBc2 | 31-120 GAG 204-297 GAG 330-492 | 62 |
| ΔV2ΔV1/2ΔV3ΔC5 | HXBc2 | 83-127 GAG 195-297 GAG 330-492 | 62 |
| ΔV2ΔV1/2ΔV3ΔC5 | HXBc2 | 83-127 GAG 195-302 GAG 325-492 | 62 |

* The ΔV1/2V3 and ΔV1/2V3ΔC5 constructs were chimeras of strains BH10 and HXBc2.

* The V1/2* and V3 constructs were deglycosylated with various deglycosylating enzymes (singly or in combination) in 0.5 M NaCl, 100 mM sodium acetate, pH 5.7, for 10 h at 37 °C. Endoglycosidase D was used at a concentration of 0.1 unit/ml, endoglycosidase F at
Results and Discussion

To address the many problems associated with the crystallization of HIV-1 gp120, we exploited the modularity of the macromolecular surface using tactics that target protein modification and conformational restriction (Table III). Several of these tactics contain novel features and are detailed here.

**Variant Constructs of the gp120 Protein**—Variants of gp120 were developed through an iterative cycle which strove to eliminate heterogeneity. The cycle involved recombinant production of gp120 variants, deglycosylation, and then assessment of heterogeneity and flexibility by examinations of glycosylation status, monoclonal antibody binding, and protease sensitivity, leading to the design of new constructs. For example, protease digestion monitored by PAGE indicated susceptability at the C terminus, and a form with 15–20 residues removed by carboxypeptidase Y retained CD4 binding activity. A homogeneous product was difficult to make by this method, and primer-based polymerase chain reaction mutagenesis and recombinant expression were used to generate a homogeneous gp120 derivative with a 19-residue C-terminal deletion. At the N terminus, sequencing of the initial constructs showed the expected signal cleavage at +51, with four additional amino acids, Gly-Ala-Arg-Ser, added from the signal peptide (a consequence of different processing of the cloning vector signal peptide with gp120).

**Conformational heterogeneity**—Conformation restriction with protein ligands such as CD4 and Fab from conformationally sensitive monoclonal antibodies 

\[
\left(\frac{p_{\text{ave}}}{p_{\text{ave}}} - 1\right)
\]

**Crystallization**—The vapor-diffusion hanging-drop technique was used for most crystallizations, screening solutions were recalibrated to account for the observed 20 °C precipitation point and another full screen at 4 °C was set up. If no crystals were observed, the Crystal Screen II (Hampton Research) was set up at 20 °C.

**Optimization**—In addition to the standard single variable optimization of crystallization conditions, a factorial-like procedure was used to determine if small amounts of different additives increased crystal quality. Type E crystals were grown from the following conditions: protein (ΔS2A V1/2 ΔV5ΔC5) gp120, two-domain CD4 (D1D2), Fab 17b purified as a ternary complex on the Superdex S-200; droplet (0.5 μl of protein solution consisting of ~10 mg/ml protein in gel filtration buffer + 0.4 μl droplet mixture containing 0.1 M sodium citrate, 0.02 M Na-Hepes, 10% isopropyl alcohol, 10.5% PEG 5000 monomethylether (Fluka), 0.0075% SeaPrep-agarose (FMC BioProducts), pH 6.4; Reservoir (0.25 M NaCl, 0.1 M sodium citrate, 0.02 M Hepes, 10% isopropyl alcohol, 10.5% PEG 5000 monomethylether, pH 6.4). The droplet mixture was kept at 37 °C to ensure the agarose solubility, and the crystallization setup at room temperature. Clumps of crystals appeared within 2 weeks of incubation at 20 °C and grew for several months to maximal size.

**X-ray Diffraction Characterization**—All data were collected at beamline X4A of the National Synchrotron Light Source, Brookhaven National Laboratory. The type E crystals were cross-linked with the vapor diffusion technique of Lighty (48) by placing a crystallization bridge (Hampton Research) with a 25-μL sitting droplet of 1% gluteraldehyde (Sigma) in the reservoir of a standard hanging-drop vapor diffusion crystallization setup for 1 h at room temperature. The cross-linked crystal was washed with stabilizer (reservoir solution with only 50 μM NaCl containing 10% ethylene glycol. After approximately 24 h, the external liquid surrounding the crystal was replaced with paratone-N (Exxon), the crystal mounted in an ethylene loop (Hampton Research) (49), and flash-cooled in the nitrogen stream of a cryostat (details are provided in (50)). Oscillation data were processed with DENZO (51) and scaled with SCALEPACK (51).

Results and discussion

To address the many problems associated with the crystallization of HIV-1 gp120, we exploited the modularity of the macromolecular surface using tactics that target protein modification and conformational restriction (Table III). Several of these tactics contain novel features and are detailed here.

**Variant Constructs of the gp120 Protein**—Variants of gp120 were developed through an iterative cycle which strove to eliminate heterogeneity. The cycle involved recombinant production of gp120 variants, deglycosylation, and then assessment of heterogeneity and flexibility by examinations of glycosylation status, monoclonal antibody binding, and protease sensitivity, leading to the design of new constructs. For example, protease digestion monitored by PAGE indicated susceptibility at the C terminus, and a form with 15–20 residues removed by carboxypeptidase Y retained CD4 binding activity. A homogeneous product was difficult to make by this method, and primer-based polymerase chain reaction mutagenesis and recombinant expression were used to generate a homogeneous gp120 derivative with a 19-residue C-terminal deletion. At the N terminus, sequencing of the initial constructs showed the expected signal cleavage at +51, with four additional amino acids, Gly-Ala-Arg-Ser, added from the signal peptide (a consequence of different processing of the cloning vector signal peptide with gp120). Protease digestion gave a product at +40, indicating flexibility in the N terminus. Progressive genetic truncation
and biochemical analysis identified +83 as a variant that was recognized by conformation-dependent gp120 ligands, whereas +94 exhibited some conformational disruption (46). Thus much of the apparently flexible region at the N terminus of gp120 could be removed without disrupting the global conformation of the protein.

To further reduce flexibility, variable loops, V1, V2, and V3, were deleted and replaced with shorter segments, as reported earlier (52, 53). Little effect was found on CD4 binding activity (47, 53). Three constructs were made which contained deletions of the V1, V2, and V3 loops (Table II). In the ΔV1/2ΔV3 construct, the entire base and stem of the variable loops V1, V2, and V3 were excised. In the ΔV1/2ΔV3 protein, the conserved stem of the V1/V2 stem-loop structure was retained, restoring the CD4-induced antibody epitopes in the presence of soluble CD4. In the ΔV1/2ΔV3* protein, the base of the V3 loop was retained as well, fully restoring CD4-induced antibody epitopes, even in the absence of soluble CD4.

Deglycosylated Forms of gp120—The asparagine-linked carbohydrate on the gp120 glycoprotein produced in Drosophila cells was analyzed. Dionex chromatography showed that the carbohydrate on this protein consisted of (N-acetylgalactosamine)2 (fucose)4-acetylglucosamine, and four with N-acetylgalactosamine, and four with fucose and +94 exhibited some conformational disruption (46). Thus much of the apparently flexible region at the N terminus of gp120 could be removed without disrupting the global conformation of the protein.

To further reduce flexibility, variable loops, V1, V2, and V3, were deleted and replaced with shorter segments, as reported earlier (52, 53). Little effect was found on CD4 binding activity (47, 53). Three constructs were made which contained deletions of the V1, V2, and V3 loops (Table II). In the ΔV1/2ΔV3 construct, the entire base and stem of the variable loops V1, V2, and V3 were excised. In the ΔV1/2ΔV3 protein, the conserved stem of the V1/V2 stem-loop structure was retained, restoring the CD4-induced antibody epitopes in the presence of soluble CD4. In the ΔV1/2ΔV3* protein, the base of the V3 loop was retained as well, fully restoring CD4-induced antibody epitopes, even in the absence of soluble CD4.

Deglycosylation with enzymes such as glycopeptidase F (or endoglycosidases D or H, leaving only a single N-acetylgalactosamine residue and, potentially, a 1–6 fucose attached to any of the glycosylated asparagine residues, appeared to leave the protein intact as judged by a panel of conformationally sensitive monoclonal antibodies (47). Digestion of full-length constructs with endoglycosidase H, which has specificity for oligosaccharides with 5–9 mannose residues, removed roughly 60% of the carbohydrate, and addition of endoglycosidase D, which cleaves oligosaccharides with 3 or 4 mannose residues, removed up to 90% of the carbohydrate. For the variable loop-deleted constructs, all mannose residues were removed with the endoglycosidase D/H combination as judged by carbohydrate analysis and by the inability of concanavalin A to bind to the deglycosylated protein. Mass spectroscopy of the deglycosylated ΔS2ΔV1/2ΔV3ΔC5 gp120 showed a molecular mass of 39,000 ± 50 Da, consistent with a mass of 35.4 kDa for the protein (based on the DNA sequence) and 3.6 kDa for the remaining carbohydrate. Carbohydrate analysis showed only fucose and N-acetylgalactosamine sugars to be present, in a ratio of 1.3 ± 0.2, respectively. These results suggest that, of the 18 potential asparagine glycosylation sites in the ΔS2ΔV1/2ΔV3ΔC5 gp120, five are unused, nine are modified with N-acetylgalactosamine, and four with N-acetylgalactosamine (1–6-fucose).

Complexes with gp120 Ligands—Protein ligands, CD4, and the Fab fragments of monoclonal antibodies, were used in an attempt to reduce mobility in the overall surface of the protein and, hence, in the potential crystal lattice. This was complicated by the internal mobility of these ligands: CD4 has a flexible juncture between the second and third extracellular domains (54), and Fabs have a conformationally mobile "elbow bend" between their variable and constant domains (55). For CD4, we used a construct containing the N-terminal two domains (1–182), for which we had previous success in structure determination (27). Fabs of the monoclonal antibodies were screened individually, even though combinations of Fabs were possible.

Initial trials with the Fab 178.1, which recognizes a linear epitope in V3 of both free and CD4 bound gp120 (44), gave only crystalline precipitates at best. We also tested the Fab of the anti-CD4 antibody L71, which recognizes the CDR3-like loop in domain D1 (45), but had difficulties preparing ternary complexes, probably due to a destabilization of the CD4-gp120 interaction. Subsequently, we focused on gp120-directed antibodies with discontinuous epitopes, which were more likely to recognize conformationally rigid portions of gp120. Complexes of gp120 proteins with Fabs of C11, which recognizes an epitope spanning C1 and C5 (42), and F105, whose epitope lies within C2, C3, C4, and C5 (overlapping the CD4 binding site) (43) gave only poor crystals (Table IV). We had greater success with 17b, which not only recognizes a discontinuous epitope but discriminates between different conformational states of gp120 (36). The Fab of 17b did not bind the initial gp120 constructs, requiring the restoration of the stem of the V1/V2 loop (constructs ΔV1/2ΔV3 or ΔV1/2ΔV3*).

Crystallization—We screened 18 different combinations of gp120 variants and ligands (Table IV), using a limited factorial based crystallization screen. Factorial screening was originally devised as a method for deducing the essential crystallization factors from combinations of different conditions (1). The empirical observation, however, that most crystallizable macromolecules are able to crystallize from a limited set of common conditions, has validated an entirely different process: crystallization screening with a small but diverse collection of fixed conditions (2). A high probability of success has been reported with as few as 6 different conditions at 4 different concentrations (56), and commercial kits are available with 50–100 conditions (Hampton Research).

In conjunction with the limited crystallization screen, small volume droplets were used, typically 0.5 μl of protein per crystallization trial. With small volumes, 1–2 mg of protein was sufficient to evaluate each gp120 crystallization variant. Smaller volumes were also more efficient at nucleation than larger droplets, perhaps due to higher surface tension effects which may result in a greater range of precipitant concentrations for each droplet to sample. Indeed, droplets that were "spread-out" also showed enhanced nucleation. This explanation may also account for the well known observation that crystals frequently nucleate from the edges of crystallization droplets.

The initial crystallization screens produced six different types of crystals (Fig. 1, Table V). For crystal types A-D, extensive optimization was unable to produce single crystals large enough to be characterized. For crystal types E and F, single crystals of needle morphology could be grown. The growth of single crystals of type E, however, required the addition of agarose, which was identified during optimization by the additive screening process. Trials with a variety of agaroses found that SeaPrep, with a gelling point near room temperature, gave the best results. Despite considerable effort, further crystallization optimization failed to produce large single crystals, and the best typical crystals were rods with a cross-section of only 30 × 40 μm. A closely related crystallization variant, which retained 10 additional amino acids in the stem of the V3 loop, failed to crystallize (Table IV).

Characteristics of gp120 Crystals—Single crystals of type E and F were analyzed for diffraction in capillary mounts. Only type E crystals showed diffraction. The needle axis of type E crystals proved to coincide with the a axis, and the rhombohedral cross-section perpendicular to the needle axis proved to be bounded by faces of the form (0 1 1). These could be distinguished from type F crystals, where the cross-section was hexagonal.
Conclusions—Our success with gp120 demonstrates the power of variational crystallization. We have derived equations that quantify the effect of this strategy on the overall probability of crystallization and have calculated the corresponding probability enhancements for several of the biochemical and molecular biological manipulations employed in this study. As can be seen (Table III), the probability of crystallization can be strongly influenced by reducing molecular surface heterogeneity. The influence of using multiple variants is more difficult to quantify since it depends on the individual probability of crystallization for each variant. Nonetheless, our theoretical analysis shows that the effect of multiple variants is greatest for proteins least likely to crystallize.

While the variational approach with gp120 did involve extensive effort, this was primarily a consequence of the difficulty in producing the gp120 glycoprotein, which involved expression levels of only a few mg of gp120 per liter of eukaryotic cell culture. While future advances in molecular biology will no doubt make such projects less arduous, if proteins are expressed bacterially, present day recombinant techniques coupled to affinity or “tag” purifications make the generation of variants straightforward. A recent example, involving the generation of 11 different variants in the crystallization of an anagical. Gel electrophoresis of type E crystals demonstrated that they contained all the elements of the ternary complex: gp120, D1D2, and Fab 17b (Fig. 2).

We were unable to flash-cool the type E crystals with standard cryoprotectants. Satisfactory results were found with a procedure that (i) fortified the crystals with vapor-diffusion glutaraldehyde cross-linking (48), (ii) permeated the crystals with 10% ethylene glycol, and (iii) used an immiscible oil, paratone-N, to replace the external solution around the crystals prior to flash-cooling (50). Cryopreserved crystals diffracted to Bragg spacings of better than 2 Å, although the diffraction was anisotropic, with higher mosaicity along the 88 Å b-axis.

Type E crystals were orthorhombic, space group P222, with unit cell parameters, a = 71.25 Å, b = 88.11 Å, and c = 196.44 Å (α = β = γ = 90°). Solvent content analysis yielded a solvent content of 58% for one ternary complex in the crystallization asymmetric unit (assuming partial specific volumes of 0.73 for protein and 0.65 for carbohydrate and the observed total molecular mass of 108.3 kDa for the complex of which 3.6 kDa is carbohydrate). Diffraction data have been collected to a limit of 2.2-Å spacings (Table VI).

We were unable to flash-cool the type E crystals with standard cryoprotectants. Satisfactory results were found with a procedure that (i) fortified the crystals with vapor-diffusion glutaraldehyde cross-linking (48), (ii) permeated the crystals with 10% ethylene glycol, and (iii) used an immiscible oil, paratone-N, to replace the external solution around the crystals prior to flash-cooling (50). Cryopreserved crystals diffracted to Bragg spacings of better than 2 Å, although the diffraction
The failure of proteins to crystallize is rarely reported in the literature, and our own results comprise too small a sample to be statistically meaningful. Nonetheless, we note that for gp120 the algorithms predict that crystals are most probable with deglycosylation, variable loop removal, and addition of an ordered protein ligand. Consistent with prediction, for the 6 crystallization variants that did have all of these modifications, three (or 50%) produced crystals, whereas for the 12 variants that did not have these modifications, no crystals (0%) were grown. In addition, theory predicts that well ordered crystals are most probable when the overall probability of crystallization is highest; Table IV shows that the crystallization variant that produced the only well ordered crystals appeared to have the greatest probability of crystallization, producing three different crystal forms whereas the best of the other variants only produced one form.

The crystallization literature is replete with examples of protein manipulation, from proteolytic digestion, to variation in solvating detergent, to screening of DNA oligonucleotides (38). What distinguishes our efforts is the derivation of a theoretical foundation, which allows the probabilistic assessment of the most effective crystallization approach. Because of the conformational complexity of gp120, we focused on surface modification, to eliminate heterogeneity and to present new crystallization variants, coupled to a limited screen of crystallization conditions. The types of crystallization problems encountered in gp120 (Table III) are not so different from many of the typical problems facing present day crystallographers; both from a theoretical or from a practical perspective, the strategy of probability analysis coupled to variational crystallization may be broadly applicable.

Subsequent to the submission of this manuscript, the structure determination of type E crystals was reported (63).

Acknowledgments—We thank Mary Ann Gawinowicz and Andrew Pound for N-terminal sequencing and carbohydrate analysis, Craig Ogata for beamline assistance, and past and present members of the Hendrickson group, especially Arno Pahler for his maxim, "The most important variable in a protein crystallization is the protein itself." We thank the Biopharmaceuticals Division of SmithKline for contributions to the expression, production, and purification of gp120 and CD4 proteins, particularly M. Strohsacker and D. Kokolis. X-ray diffraction data were collected at beamline X4A, National Synchrotron Light Source, Brookhaven National Laboratory.
Probability Analysis of Variational Crystallization and Its Application to gp120, The Exterior Envelope Glycoprotein of Type 1 Human Immunodeficiency Virus (HIV-1)

Peter D. Kwong, Richard Wyatt, Elizabeth Desjardins, James Robinson, Jeffrey S. Culp, Brian D. Hellmig, Raymond W. Sweet, Joseph Sodroski and Wayne A. Hendrickson

J. Biol. Chem. 1999, 274:4115-4123.
doi: 10.1074/jbc.274.7.4115

Access the most updated version of this article at http://www.jbc.org/content/274/7/4115

Alerts:
- When this article is cited
- When a correction for this article is posted

Click here to choose from all of JBC’s e-mail alerts

This article cites 59 references, 17 of which can be accessed free at http://www.jbc.org/content/274/7/4115.full.html#ref-list-1