A Visual Language for Protein Design
Appendices: Details on Protein Glyphs and Drawing Rules

Robert Sidney Cox III, James Alastair McLaughlin, Raik Grünberg,
Jacob Beal, Anil Wipat, Herbert M Sauro

Appendix A: Protein Glyphs

- **Protein region, unspecified.** A protein region of unspecified function and structure. This denotes a generic region of protein, which is not known to contain a structured domain, such as an unstructured linker sequence. This glyph is also referred to as a protein backbone, since other glyphs can attach to it.

- **Protein region, omitted** The dashed line is provided for omitting a region of the protein backbone, in cases where the relative scale of the protein domains should be shown but portions of the protein are omitted from the diagram to maintain the relative scale.

- **Membrane region.** The membrane region glyph can be used to modify either unspecified or structured protein regions. The glyph is placed along either backbone, interrupting it with a membrane glyph. Membrane regions include both trans-membrane regions and membrane anchors into or across a plasma membrane or organelle (biological context or other diagrams must be used to distinguish between types of membrane regions). The site glyphs below cannot be put on top of a membrane glyph to indicate a functional modification in a membrane region. Instead, site glyphs can be placed vertically aligned with a membrane glyph on a structured region backbone.

- **Protein region, structured.** A protein region with specific function or structure, such as a protein domain. Structured regions can be re-sized horizontally to have variable lengths. Structured regions can include multiple protein domains, and generally protein fusions would be represented as two or more structured regions. The structured region can be decorated with other site glyphs on the top or bottom backbone. This region can also be used to represent any function not covered by other glyphs.

- **Catalytic site**. A protein region including one or more enzyme active sites. This glyph is meant to be flexible, so it could be used to label anything from an entire active site, down to as small as a single residue that participates in an enzyme active site. Multiple glyphs can be used to show the relative positions of particular active site residues. The label letters and colors can be used to indicate different substrates and different sites contributing to the same active site.

- **Binding site, non-covalent**. A general ligand binding site or dimerization domain. This will often represent a non-covalent contact between a protein site and another protein or peptide, a small molecule binding site, or binding to another macromolecule such as DNA. We do not proscribe how to draw binding interactions, just the presence of the binding site and a label. Binding sites can affect protein conformation and enzyme activity.

- **Protease cleavage site.** A polypeptide region that directs a protease cleavage of the peptide chain. The cleavage could be modulated by protein conformation, or other protein sites such as catalytic sites or covalent modifications.

- **Covalent modification.** A site for a covalent attachment to a protein such as phosphorylation, methylation, etc., with type indicated by a letter. Covalent sites can affect protein conformation and enzyme activity.
- **Localization signal, cleaved.** A targeting polypeptide that encodes the localization of the protein. The signal peptide is then cleaved, and therefore the signal can not be used again.
- **Localization signal, retained.** A targeting polypeptide that encodes the localization of the protein. Since the signal is retained after transport, it is possible for this signal to be used repeatedly.
- **Degradation signal.** A sequence that directs protein degradation. For example, in bacteria this could be a recognition sequence for a protein degrading enzyme, and in eukaryotes this could be a ubiquitination site.
- **Biochemical tag.** A sequence that is useful for biochemical manipulation, detection or characterization. Examples include the His tag for protein purification and the FLAG tag for antibody recognition.

Appendix B: Glyph Drawing Rules

Any implementation of Protein Language or extension thereof should conform to the following rules:

Protein Backbone

The protein backbone is the line on which the protein design glyphs are drawn. In some cases (e.g., structured protein regions) the line is not shown in the design, in other cases (e.g., protein sites) the line is shown with the glyphs on top. We do not restrict the shape of the line, though it should not cross itself. The protein line is drawn with a default 3pt thickness. Other linewidths are automatically considered to be labels. Multiple amino acid chains (such as for protein complexes) should be drawn as distinct protein lines.

Protein Regions

Regions are amino acid chains that can be of variable size, typically more than ten amino acids. Regions are drawn as replacements for sections of the backbone line. Any region glyph is scaleable in the dimension that is parallel to the backbone line, but may not be scaled in the other (i.e., the dimension normal to the backbone line). Regions can be expanded horizontally provide information about different sizes of amino acid regions, or to accommodate labels. When drawn on a curve or irregular line, the region glyph should lie along a straight line which connects the two endpoints of the backbone line (thus the backbone line is replaced with a straight line where there is a region glyph). Regions may not be drawn overlapping each other.

Protein Sites

Sites are generally smaller than regions, typically one to fifty amino acid residues. Sites are drawn as glyphs on top of the backbone line, centered, with the backbone line displayed below. Sites can also be drawn onto the structured region glyph (see Examples for implementation details). Site glyphs are not scaleable, so they are always the same size relative to the protein line. Site glyphs cannot be drawn on top of each other or overlapping. Sites may be drawn with a single character label (additional labels can be shown in a separate "labels" layer of a diagram). Sites can restrict to a specific set of characters such as lower case roman, upper case roman, roman numerals, etc. to the visual appeal of the glyphs. Protein Designer restrictions are: lower case in squares, upper case in diamonds and circles.