

Identification and insertion of 3-carbon bridges in protein disulfide bonds: a computational approach

Mire Zloh¹, Sunil Shaunak², Siby Balan³ & Steve Brocchini³

¹Department of Pharmaceutical and Biological Chemistry, University of London, 29/39 Brunswick Square, London WC1N 1AX, UK. ²Faculty of Medicine, Imperial College London, Hammersmith Hospital, Ducane Road, London W12 0NN, UK. ³Department of Pharmaceutics, The School of Pharmacy, University of London, 29/39 Brunswick Square, London WC1N 1AX, UK. Correspondence should be addressed to M.Z. (mire.zloh@pharmacy.ac.uk).

Published online 26 April 2007; doi:10.1038/nprot.2007.119

More than 42,000 3D structures of proteins are available on the Internet. We have shown that the chemical insertion of a 3-carbon bridge across the native disulfide bond of a protein or peptide can enable the site-specific conjugation of PEG to the protein without a loss of its structure or function. For success, it is necessary to select an appropriate and accessible disulfide bond in the protein for this chemical modification. We describe how to use public protein databases and molecular modeling programs to select a protein rationally and to identify the optimum disulfide bond for experimental studies. Our computational approach can substantially reduce the time required for the laboratory-based chemical modification. Identification of solvent-accessible disulfides using published structural information takes approximately 2 h. Predicting the structural effects of the disulfide-based modification can take 3 weeks.

INTRODUCTION

Disulfide bonds influence the physico-chemical and biological properties of proteins in many subtle ways¹. Free Cys are rare in proteins with disulfides and, if present, are usually buried. Small proteins contain more disulfide bonds than large proteins because the former need to compensate for the low number of hydrophobic contacts². Most disulfide bonds are intra-chain rather than inter-chain³. It is generally accepted that the removal, modification or addition of disulfides in proteins using recombinant technologies will result in a loss of function⁴. Their simple chemical modification can also impair function^{5,6}, albeit with notable exceptions^{7,8}.

We have recently described a method for the site-specific conjugation of PEG to proteins, peptides and Ab fragments by the chemical modification of their accessible disulfide bonds. It involves the insertion of a 3-carbon bridge between the two sulfurs of a native disulfide bond^{9–11}. Our strategy relies upon the reduction of an accessible disulfide bond followed by *bis*-alkylation to insert the 3-carbon bridge to which PEG has been covalently attached. The bridge reconnects the two sulfur atoms from the original disulfide bond (Fig. 1). Despite this local change in structure, the protein's tertiary structure is not altered and biological activity is preserved^{9,10}.

Exploitation of the site-specificity and the chemical efficiency of the two sulfurs from an accessible disulfide with a single reagent contrasts with most other types of chemical modification. For the most part, each sulfur undergoes separate and independent chemical modification. In general, disulfides are located either in the buried regions of the protein's folded structure or on its solvent-accessible surface. As our approach focuses on the modification of the disulfides that are accessible to solvent¹¹, it is important to identify them at the start of the study. These accessible disulfides are more likely to contribute to the stability of a protein than to its structure or function¹². As a result, it is possible to re-bridge some of them without altering the protein's structure or function. In the course of our studies, we have used a variety of computational tools to identify these 'bridgeable' disulfides and to model the structural consequences of inserting a 3-carbon bridge. Using therapeutically relevant proteins, we have established that it is possible to insert a 3-carbon bridge across a native disulfide bond and retain the protein's tertiary structure and biological activity^{9,10}.

Publicly available protein databases and molecular modeling programs can be easily accessed via the Internet and used to examine a protein's structure and estimate the relative surface accessibility of each protein disulfide bond. With this information, it is possible to calculate the structural effects of inserting a 3-carbon bridge into each of a protein's disulfide bonds^{9,10}. As a native disulfide has to be reduced to free the Cys sulfurs for the chemical reaction, our computational approach makes it possible to estimate the propensity of a protein to maintain its tertiary structure when its disulfide bonds have been reduced. The effect of inserting a 3-carbon bridge across the original disulfide bond on the protein's tertiary structure can then be defined (Fig. 1). Although we have shown that the 3-carbon bridge can be used to conjugate PEG efficiently in aqueous solution^{10,13}, insertion of the bridge also offers the potential to enable the site-specific conjugation of pharmacological and imaging agents to the protein^{14–16}.

The Protein Data Bank (PDB) at <http://www.pdb.org> contains more than 40,000 publicly available 3D structures of proteins and peptides whose structures have been experimentally solved. Using

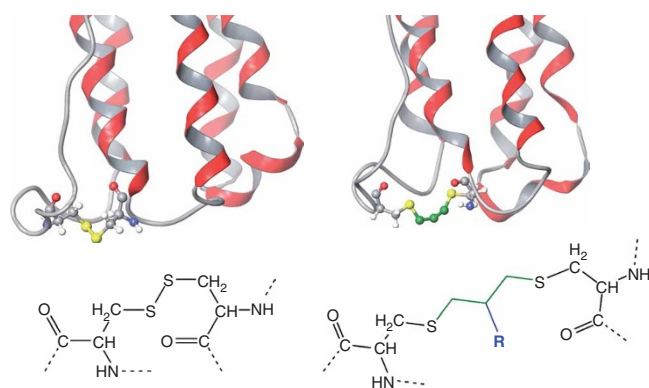


Figure 1 | Native (yellow) and 3-carbon-bridged disulfide bond (green) of interferon α -2a. R represents the functionality attached to the 3-carbon bridge.



this structural information, other publicly available molecular graphics programs can be used to find the disulfide bonds and to determine their suitability for modification with a 3-carbon bridge. Our experience indicates that this protocol can be used to determine whether a protein is suitable for modification before embarking on laboratory-based experiments. Over a period of time, the original model can also be refined using the experimental data generated. In addition, there are more than 350,000 entries for human proteins in the protein sequence data bank whose sequences are known but for which there is insufficient tertiary structural information (Entrez Protein, <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Protein>). If the complete structure of a protein is not known, then homology modeling or *ab initio* protein modeling can be used to predict its 3D structure¹⁷. Depending on the amount of structural information available, these particular modeling strategies may require additional specialist expertise.

Strategy layout

This protocol is divided into two parts. Part 1 describes the strategy for identifying whether a protein is suitable for the insertion of a 3-carbon bridge across its disulfides. It aims to locate the position of the disulfide bonds in a protein and can be undertaken by a scientist without molecular modeling experience, using free molecular graphics software packages. Slight modifications of the procedure described can also be used to examine the position of other residues and their accessibility to solvent.

Part 2 of the protocol describes a procedure for predicting potentially adverse changes in a protein's structure as a result of inserting a 3-carbon bridge. Publicly and commercially available graphics and modeling tools can also be used to perform these prediction studies. It is useful to have some expertise in modeling studies when carrying out this part of the protocol. The results obtained with different modeling tools should be similar if the strategies described are followed. Although our protocol was developed and tested in detail for the site-specific PEGylation of interferon α -2b, we have shown that it can be modified and applied to other proteins and peptides¹³.

To provide a broader perspective, we provide a brief overview of the two parts of the protocol.

Strategy for part 1 of the procedure—identifying the presence and location of disulfides

The suggested procedure for identifying whether a protein and its disulfides are suitable for the insertion of a 3-carbon bridge is shown as a flowchart in **Figure 2**.

Although the PDB can be searched for the protein structure of interest, it is important to note that the structure obtained may be incomplete because of the absence of flexible loops and domains. In such cases, some of the disulfide bonds may not be shown.

```

1 maltfallva llvlsckssc svqcdipqth slgsrrtlml laqmrrislf elkdrhdhg
61 fpqeeefgnqf qkaetipvlh emiqqifnlf stkdssaawd etlldkfyte lyqqindlea
121 cviggvgvte tplmkedsil avrkyfgrit lylkekkysep cawevvraei mrsfslstnl
181 qeslrskc
    
```

Figure 3 | The sequence of interferon α -2a (LOCUS number: NP_000596) from the Entrez Protein database (<http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi>). Cys residues (c) are shown in red. The signal sequence is shown in blue. There are two Cys residues in the signal sequence. They are shown in blue.

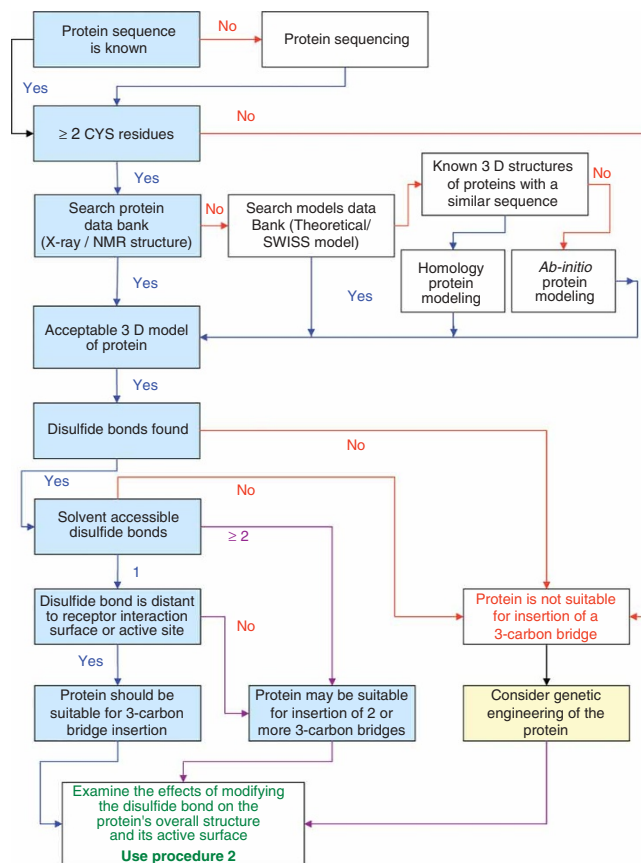


Figure 2 | Strategy to identify a protein's accessible disulfide bonds, and suitability for modification with a 3-carbon bridge.

The general procedure that should be used is as follows:

- To account for all of the Cys in a protein, we recommend that the protein's sequence is used. When a sequence is found in the PDB, it may be associated with literature citations that describe the protein—for example, active site or physico-chemical properties. It is useful to save the LOCUS number and the sequence for future reference (**Fig. 3**).
- Determining whether a disulfide bond is present and whether it could be a target for the insertion of a 3-carbon bridge requires additional structural information. The PDB then needs to be searched to determine whether the experimentally determined 3D structure of the protein has been published. If the structure is available, the atomic coordinates of the protein should be saved in a 'PDB' file format. This format is described on the web page http://www.pdb.org/pdb/file_formats/pdb/pdbguide2.2/guide2.2_frame.html.

If the structure is available, it is possible to determine (i) the presence of disulfide bonds, (ii) the location of the disulfides [hydrophobic (interior) versus solvent accessible (surface)] and (iii) the position of the accessible disulfide bonds relative to the protein's receptor/substrate binding surface(s). At this point, it is important to check the protein's structure for the correct numbering and sequence completeness. When 3D data are available for only some sections of the protein, the missing structural features can be built into the protein using a combination of homology modeling and molecular dynamics simulations (**Fig. 4**).



When an experimentally determined structure cannot be found or is incomplete, the following two databases can be searched for theoretically modeled structures: Protein Model Database (<http://a.caspr.it/PMDB>) and SwissModel repository (<http://swissmodel.expasy.org/repository>). Although these two databases will not provide an exact structure of the protein, they can provide an insight into the protein's topology. If a model structure cannot be found, it is still possible to define the theoretical structure of the protein using homology modeling or *ab initio* protein modeling approaches¹⁷. The effective use of these approaches requires considerable specialist expertise and, as such, is beyond the scope of this protocol.

There are more than 30,000 entries in the PDB with Cys residues and 1,160 entries when the keyword 'disulfide' is used. As a 'disulfide' is not always defined explicitly in the PDB file, the protein structure obtained should be tested to determine whether any disulfide bonds are present. This can be achieved by visualizing the Cys residues. If the Cys residues are in close proximity to each other (2–3 Å) in the folded structure, they are likely to form a disulfide bond. A particularly useful tool that enables disulfides to be determined with accuracy is the software Disulfide by Design, where paired Cys residues can be identified by the geometric constraints on the sulfur atoms in disulfide bonds (Table 1). For example, in disulfides of known protein structures, the torsion angle formed by the Cβ–Sγ–Sγ–Cβ bonds and the rotation about the Sγ–Sγ bond is bimodal with sharp peaks at +100° and –80°. The distribution of the Cα–Cβ–Sγ angles in known disulfides has a peak of approximately 115° and a range of 105° to 125° (ref. 18).

Many disulfide bonds are not accessible to solvent because they are buried in the protein's hydrophobic regions. Consequently, it is difficult for reagents to reach them under non-denaturing conditions. In addition, reducing these disulfides can significantly impair the protein's ability to refold to its original tertiary structure¹⁹. Therefore, a lack of solvent-accessible disulfides means that the protein is unlikely to be suitable for the insertion of a 3-carbon bridge. Disulfides that are close to or on the protein's surface are more likely to be available for chemical reactions under non-denaturing conditions and with minimal stoichiometries of reagents. The solvent-accessible surface area (SASA) can be defined and used to estimate the accessibility of these disulfides. MOLMOL²⁰ is an example of a software tool that can be used to examine the accessibility of a disulfide to chemical reagents. For example, Table 2 shows the MOLMOL-calculated SASAs of the Cys residues of the proteins that we selected for the insertion of 3-carbon bridges after the reduction of disulfide bonds. Please note that the results obtained using this approach should be confirmed by visual inspection using a molecular graphics package such as Jmol or Rasmol. It is also important to remember that 3D models provide only a static picture of the protein; a protein and its side chains are constantly in motion, with partially hidden sulfur atoms being

TABLE 1 | Summary of the geometrical parameters of disulfide bonds detected in the PDB¹⁸.

Parameter	Value	Range
Distance between S atoms in two Cys	—	2–3 Å
Angle Cα–Cβ–Sγ	115°	±10°
Torsional angle Cβ–Sγ–Sγ–Cβ	+100° or –80°	±10°

a

ATCM	171	N	ILE	24	63.505	-31.382	-10.621	1.00	59.84	N
ATCM	172	CA	ILE	24	62.976	-32.691	-10.984	1.00	65.48	C
ATCM	173	C	ILE	24	62.126	-32.524	-12.234	1.00	75.22	C
ATCM	174	O	ILE	24	62.239	-33.303	-13.178	1.00	73.90	O
ATCM	175	CB	ILE	24	62.119	-33.309	-9.858	1.00	64.00	C
ATCM	176	CG1	ILE	24	63.019	-33.778	-8.713	1.00	62.30	C
ATCM	177	CG2	ILE	24	61.278	-34.462	-10.392	1.00	64.54	C
ATCM	178	CD1	ILE	24	62.298	-34.556	-7.647	1.00	58.42	C
ATCM	179	N	LEU	39	49.861	-27.515	-11.417	1.00	76.50	N
ATCM	180	CA	LEU	39	50.067	-27.949	-10.028	1.00	69.16	C
ATCM	181	C	LEU	39	50.795	-26.838	-9.249	1.00	65.31	C
ATCM	182	O	LEU	39	51.469	-27.106	-8.245	1.00	62.76	O
ATCM	183	CB	LEU	39	50.929	-29.217	-9.995	1.00	62.94	C
ATCM	184	CG	LEU	39	50.399	-30.525	-10.564	1.00	51.12	C
ATCM	185	CD1	LEU	39	51.569	-31.393	-10.944	1.00	50.74	C
ATCM	186	CD2	LEU	39	49.543	-31.215	-9.548	1.00	46.81	C
ATCM	187	N	ASP	40	50.667	-25.607	-9.744	1.00	60.43	N
ATCM	188	CA	ASP	40	51.306	-24.434	-9.167	1.00	59.22	C
ATCM	189	C	ASP	40	50.820	-24.107	-7.776	1.00	58.67	C
ATCM	190	O	ASP	40	51.593	-23.680	-6.915	1.00	58.60	O
ATCM	191	CB	ASP	40	51.039	-23.209	-10.047	1.00	66.89	C
ATCM	192	CG	ASP	40	51.872	-23.198	-11.311	1.00	72.70	C
ATCM	193	OD1	ASP	40	53.102	-23.372	-11.205	1.00	77.43	O
ATCM	194	OD2	ASP	40	51.304	-22.988	-12.407	1.00	74.98	O

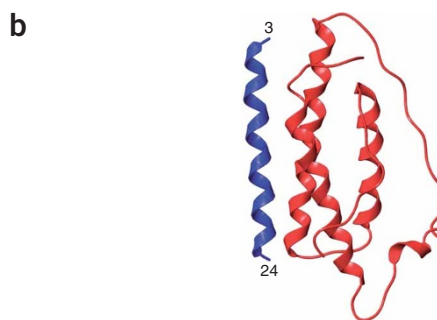


Figure 4 | Sequence and 3D model of leptin (PDB file 1AX8). It shows that part of the 3D structure of the protein is missing. (a) Text representation; (b) ribbon representation. The segment in blue is the part of the protein that is not connected to the segment in red.

constantly exposed to solvent. Some of the PDB files of proteins whose structure has been determined using NMR contain multiple conformations that could reflect the conformational mobility of the protein. With specific reference to this point, the SASAs reported in Table 2 for interferon α-2a were calculated from a set of conformations that were determined by NMR. It should also be noted that when several solvent-accessible disulfide bonds are present, it may be possible to insert more than one 3-carbon bridge.

A protein that fulfills the conditions shown in Figure 2 (i.e., has one or more accessible disulfide bonds) should be suitable for the insertion of a 3-carbon bridge across its disulfide bonds. Information on the surface of the protein that interacts with its receptor/ligand has to come from the protein's crystal structure or from biological data (e.g., mutagenesis studies) if the complex is yet to be crystallized. In general, if the disulfide bond is a part of, or close to, the protein's active binding surface, this also needs to be taken into account when considering the overall change in a protein's tertiary structure.

It now becomes possible to determine the effects of reducing the disulfide bond(s) to liberate the two Cys sulfur atoms for the insertion of a 3-carbon bridge. Chemical reduction will increase the distance between the two sulfur atoms and it may also change the local and/or global structure of the protein. This can lead to changes in the protein's tertiary structure and to significant changes in the conformation of its active surface. In addition, when two or more disulfides that are in close proximity are reduced, the possibility of 3-carbon bridge formation between sulfurs that were not originally paired as a disulfide also needs to be considered. The

TABLE 2 | The solvent-accessible surface area (SASA) of Cys residues for interferon α -2a, leptin and the Fab. PDB codes are given in brackets. Chain names are given in square brackets. The range of SASA is given for the set of conformations found in the 1ITF PDB file for interferon α -2a.

Protein	Cys-1	SASA (%)	Cys-2	SASA (%)	Classification
Interferon α -2a (1ITF)	1	8.0–23.9	98	0–3.3	Accessible
	29	0–12	138	3.2–14.4	Accessible
Leptin (1AX8)	95	17.4	146	18.4	Accessible
	23 [L]	0.3	93 [L]	0	Buried
	139 [L]	0	199 [L]	0	Buried
FAB (1CBV)	22 [H]	1	98 [H]	0	Buried
	148 [H]	0.3	203 [H]	0	Buried
	219 [L]	66.5	136 [H]	50	Accessible

procedure for determining the amount of structural change as a consequence of the chemical modification of the disulfides is described in the next section.

Strategy for part 2 of the procedure—modeling the effects of modifying the disulfide bond on the structure of the protein

The aim is to define the effect of inserting a 3-carbon bridge into each of the accessible disulfides of a protein. Evaluation of the structural behavior of the protein after modifying its disulfide bonds and PEGylation of the protein (if appropriate) can be conducted using the flowchart shown in **Figure 5**. Ideally, the disulfide into which the 3-carbon bridge will be inserted will be the most accessible disulfide bond. Its modification should result in the smallest possible change in the protein’s tertiary structure (**Fig. 5**). The first modeling study should be conducted by inserting a single 3-carbon bridge into the disulfide of a monomeric protein. In the case of multimeric proteins, our experience suggests that the best approach is to insert a single 3-carbon bridge into each of the monomers that make up the protein¹³.

As the insertion of a 3-carbon bridge requires the reduction of a disulfide, we start by determining the effect of reducing a disulfide on a protein’s structure. The molecular modeling studies described in this protocol^{9,21,22} were conducted using the integrated molecular modeling packages Maestro version 6.5 and Macromodel version 9.1. They are distributed by Schrödinger (<http://www.schrodinger.com>). The advantage of these two packages is that the graphical user interface ensures an easy setup for the calculations required, and the builder menu provides tools for the modification of disulfide bonds and for building novel structures. Although the topology and parameters for non-standard residues are automatically sorted, we recommend that the user recheck the structures and the parameters generated. Several different force fields are available and the solvent can be taken into account implicitly using the generalized Born surface area method. The additional effect of solvent collisions with the protein can be considered using stochastic (Langevin) dynamics²³. Comparative studies of molecular dynamics simulations with the explicit solvent and of stochastic dynamics simulations with the implicit solvent have shown that the results of these simulations are comparable^{9,24}.

Macromodel has some limitations when used for these kinds of studies. The current version is not suitable for running simulations in parallel, and the software is not suitable for simulation of systems with explicitly defined solvent molecules. In addition, we have encountered some problems with the number of atoms when we

tried to simulate PEGylated asparaginase using an all-atom force field. This led us to use a united atom force field.

Many free molecular dynamics software packages are available and they can also be used to achieve a similar result. However, with these software packages, it can be difficult for inexperienced users to set up the calculations for (i) new structures, (ii) novel topologies, (iii) parametrization of the atom types and (iv) connectivities between atoms. However, it is possible to select software that is suitable for the hardware available to the user. Using these programs, simulations can be run in parallel mode on cluster systems such as AMBER, NAMD or GROMACS. Simulations can also be carried out on fully solvated systems using NAMD, GROMACS or CNS-SOLVE, or using either implicit or explicit solvation methods in AMBER.

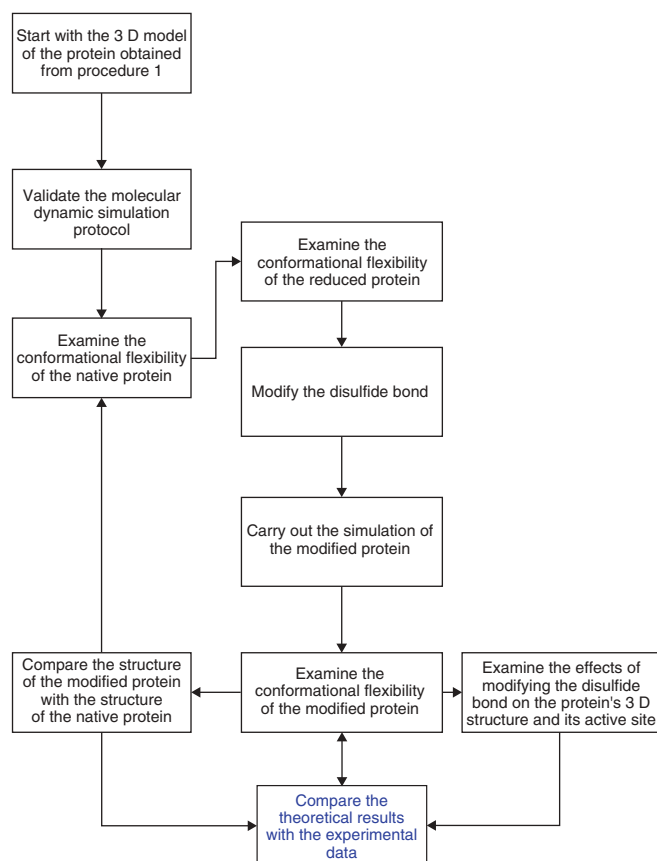


Figure 5 | Strategy for evaluating a protein’s structure after the insertion of a 3-carbon bridge across a disulfide bond.



A schematic procedure for part 2 of the protocol is now described; it is a prerequisite that a complete 3D model of the protein is available from part 1.

When using other software packages, it is necessary to validate the molecular dynamics simulation protocol. Start by running a simulation whose result is already known. Our validation procedure examines the effect of two types of simulation on the structure of a protein. The first simulation, performed using CNS-SOLVE software, is carried out with the protein fully solvated with explicit water molecules. The second is a stochastic dynamics simulation carried out with implicit solvent representation. This is implemented in Macromodel. The resulting structural simulation trajectories need to be comparable. These types of experiments will indicate whether stochastic simulations using implicit water representation (implemented in Macromodel) are suitable for modeling the protein in its (i) native, (ii) disulfide-reduced and (iii) 3-carbon-bridged forms. When other software is used, we recommend that a simulation is also conducted with interferon α -2a using the PDB entry '1ITF'.

Another important internal control of the present methodology consists of running a structural simulation of the native protein. This strategy can also be used to compare the structure of the native protein with that of the modified protein. Such an analysis provides additional information about the conformational flexibility of the native protein, which, typically, should not unfold or change its conformation at the end of the simulation.

At this time, it is also possible to run the simulation of the protein with the two sulfur atoms of a reduced disulfide bond; that is, the sulfurs are not connected. The protein's structure should not change significantly. If it does, special care may be required during the experimental work to ensure that there is minimal propensity for the protein to aggregate when the protein's disulfide is reduced. This can be achieved using additives (e.g., Arg) or by the careful selection of buffers¹¹. Knowing that special chemical conditions will be required to reduce a protein's disulfide can save experimental time and effort. In this context, it should be noted that most of the literature about the reduction of protein disulfide bonds focuses on their complete reduction; that is, including the non-accessible disulfides in the protein's interior that require protein denaturation⁶. Far less has been published about the site-selective reduction of accessible disulfide bonds.

Once it becomes clear that a disulfide bond can be broken without altering the protein's tertiary structure or that experimental conditions can be defined that continue to preserve its structure, the 3-carbon bridge can be inserted in the structure and the simulation repeated. The simulation conditions used should be the same as those used for the native protein. This step can be followed by adding PEG, if required. Insertion of the 3-carbon bridge is achieved using a molecular builder (Maestro). If PEG is to be added, its initial structure is determined using a conformational search. As our experimental data suggest that the size of the PEG does not affect the protein's biological activity^{10,13}, we use a small PEG for our modeling studies to minimize the time required for the simulation.

Structural snapshots of the modified protein from the simulation can be overlaid and the root mean square deviation (RMSD) calculated. Overlaid structures can be visualized as ribbons (i.e., traces of the backbone) and can be used as an indicator of motion in those parts of the protein that have undergone change with time. High RMSD values indicate that the protein is flexible and that its structure has changed significantly during the simulation. It is also possible to calculate the RMSD for selected regions in the protein during this analysis. The resultant trajectory can be converted into PDB files and a variety of software used to analyze the dynamic and structural properties of the protein. Software examples are MOLMOL, VMD, VEGA ZZ and gOpenMol.

The global changes in a protein's structure after its PEGylation can be evaluated by comparing the structural snapshots of the modified structure with the structural snapshots of the native protein. The structural snapshots of two different simulations can be overlaid and any global changes in secondary structure inspected visually. In addition, the RMSD values can be calculated for regions of interest (e.g., active site).

Overall, the structure of the active surface of the 3-carbon-bridged protein has to be compared with the same surface in the native protein. This comparison will determine whether significant change has occurred. In the case of PEGylation, the active surface may appear to be obscured by the PEG moiety. However, if the protein's tertiary structure is maintained, our experimental data show that the PEGylated protein continues to have biological activity^{10,13}.

MATERIALS

EQUIPMENT

- Personal computer with an Internet connection and Java-enabled web browser (see EQUIPMENT SETUP)
- Internet Explorer or Firefox software (<http://www.mozilla.com/en-US/firefox/>)
- Linux cluster with Suse 10.0 operating system and Sun Grid Engine (SGE) queuing software; as an alternative, Microsoft Windows XP/2000 or Linux Suse 10.0 or workstation or MacOS X personal computer
- Jmol (<http://jmol.sourceforge.net/>) (see EQUIPMENT SETUP)
- MOLMOL (see EQUIPMENT SETUP); as an alternative, Deep View (<http://www.expasy.org/spdbv/program/spdbv37sp5.exe>) or VMD (<http://www.ks.uiuc.edu/Development/Download/download.cgi?PackageName=VMD>) or Rasmol (<http://www.openrasmol.org/>)
- Disulfide by Design
- Schrödinger suite including Maestro and Macromodel (see EQUIPMENT SETUP); as an alternative, AMBER (<http://amber.scripps.edu/>) or CNS-SOLVE 1.1 or GROMACS (see EQUIPMENT SETUP)

EQUIPMENT SETUP

Personal computer Any modern computer that is connected to the Internet can be used for part 1 of the procedure. It requires a Java-enabled web browser. For simplicity, the setup procedure is described for a computer with a Windows operating system. If the browser is not Java enabled, the Java Runtime Environment (JRE) (Sun Microsystems; <http://java.sun.com>) has to be downloaded and used to Java-enable the browser. For some versions of Internet Explorer, select 'Tools' and then choose 'Internet Options'. In the 'Security' tab, choose the 'Custom Level' button. If the 'Java Permissions' section exists, choose Low, Medium or High safety. In the same window, 'Active Scripting' and 'Scripting of Java Applets' should then be enabled.

Jmol The Jmol software can be downloaded from <http://jmol.sourceforge.net/> by following links for the download and choosing the file for the appropriate computer platform (zip file extension). The current stable version is Jmol-11.0.RCX, where X depicts the version update. Software is under regular development, so changes in the version number can be expected. The file



'jmol-11.0.RCX.zip' can be saved in a new directory denoted 'c:\xxxx'. The term 'xxxx' is a string without spaces, because some modeling software may not support spaces in filenames. This is primarily due to historical reasons and compatibility issues with the Unix/Linux operating systems. Windows Explorer is used to find the downloaded files at the saved location, and with a right-click of the mouse, choose 'Extract All'. Specify the directory to which the file should be saved as 'xxxx'. The files are extracted and placed into the specified directory 'c:\xxxx\jmol-11.0.RCX'. The software can be started by double-clicking on the file called 'jmol.bat' or a shortcut can be created on the Desktop. This is done by right-clicking on the Desktop and selecting the 'New/Shortcut' option, and then browsing through the file system to select 'c:\xxxx\jmol-11.0.RCX\jmol.bat'. Then click OK to finalize the creation of the shortcut. Detailed instructions on using Jmol can be found on Wikipedia (http://wiki.jmol.org/index.php/Main_Page).

MOLMOL The MOLMOL software can be used as molecular graphics software to visualize the protein's structure. However, it does not depict non-protein moieties with ease. In our protocol, the use of MOLMOL to calculate the solvent-accessible surface is described. The software can be downloaded from <ftp://ftp.mol.biol.ethz.ch/software/MOLMOL/win/> as three executable files.

They can be saved in the new 'c:\xxxx\molmol' directory. Executing all three files will extract all the necessary files for MOLMOL to run. The program can be executed by double-clicking the 'molmol.exe' file in the 'c:\xxxx\molmol' directory. Alternatively, a shortcut can be created on the Desktop; that is, right-click on the Desktop and select the 'New/Shortcut' option.

Disulfide by Design software Probable disulfide bonds can be confirmed using Disulfide Bond by Design. This free software can be downloaded from the <http://www.ehscenter.org/dbd/> website after registration. The downloaded

'DbDInstall.zip' file should be decompressed, and the 'DbDInstall.exe' file is used to install the software into its default location (c:\Program Files\WSU\Disulfide by Design). Follow the on-screen instructions provided during the installation. The software is started by clicking 'Start/Programs/DisulfidebyDesign'.

Computing platforms Our preferred platform to run molecular modeling software is the Linux operating environment. Windows versions are also available for most packages, but we have not used them. The setup of the molecular modeling programs requires some knowledge of information technology and of the Linux environment, or of Windows if it is used.

Schrödinger suite including Maestro and Macromodel Macromodel is used for molecular mechanics calculations, conformational searches and stochastic and molecular dynamics simulations. It is part of the Schrödinger modeling suite and can be run through the graphical user interface Maestro. It also includes a molecule builder and is needed to insert the 3-carbon bridge. This software is commercially available and a trial version can be obtained through the <http://www.schrodinger.com> website. The software has comprehensive installation instructions, a user guide and tutorials. The procedure described here requires licenses for the use of the Maestro and Macromodel packages.

CNS-SOLVE 1.1 and GROMACS Publicly available and freely offered molecular modeling software can be found on the home pages of developer research groups such as Gromacs (<http://www.gromacs.org>) and CNS-SOLVE (<http://cns.csb.yale.edu/v1.1/>). Although this software can be used for most of the calculations described in this protocol, the lack of a graphical user interface for the calculation setup can present some difficulties for the non-expert trying to perform these calculations.

PROCEDURE

Part 1: identifying and locating disulfides ● TIMING 1–3 h depending on experience

- 1| Use a good site to browse a protein's sequence and its 3D model, such as the National Center for Biotechnology Information (NCBI) web page (<http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi>). A search can be carried out across several databases starting from this web page.
- 2| Enter a keyword or protein name into the search box. After you click on the 'GO' button, the results are presented in a table on the next screen.
- 3| Check whether a 3D model is available by observing the number of hits next to the 'Structure' link. Provided it is not zero, click on that link. The results are given as the PDB entry and its title. Although it is possible to follow the link on the PDB entry and observe the protein's 3D structure, the visualization software embedded in the PDB web pages is more powerful for this analysis. Jmol software can be used for off-line visualization as described later.
- ▲ **CRITICAL STEP** Read the description of the PDB file carefully as the appearance of the protein's name in the title does not mean that the PDB entry will have a 3D structure for the protein.
- 4| If the protein of interest is found, note the PDB entry. Then proceed to Step 10 of this procedure. If it is not found, use the search facilities of the PDB as described in Steps 5–9.
- 5| Open the PDB web page (<http://www.pdb.org>) in the web browser.
- 6| In the search box, type the name of the protein and click the 'SEARCH' button.
- 7| View the results displayed on the next page (ten results per page), which are sorted by their relevance.
- 8| If the protein of interest is not located owing to lack of hits, or if there are too many hits, narrow the search using the 'Advanced Search' feature. This allows several query criteria to be combined to widen or narrow the search. This type of query can be defined using the dropdown menu. The keyword-style query can contain a protein's name and it can be evaluated before adding the next query using the '+' button.
- 9| If many protein structures are found, it is possible to narrow the search by defining the species of the organism from which the protein was purified. For detailed help on searching and an explanation of the options available, see <http://www.pdb.org/pdbstatic/tutorials/tutorial.html>.
- 10| Once the 3D model of the protein is found, download the PDB file for an off-line analysis using Jmol. It is also possible to visualize the protein's structure online, but we will not describe this option. Click the 'Download Files' link that is located on the left-hand side of the screen for the selected protein entry. The menu will expand. Right-click on the 'PDB file' link. This will

PROTOCOL

offer a Windows menu with the 'Save Target As' option. Left-click on that option and save the file in a directory on your hard disk. The PDB files with NMR-based structures may have several different models that fulfill NMR restraints. The file should be edited and the models saved in separate files for a successful analysis.

11| Using the shortcut for Jmol, start the visualization software. Using the 'File/Open' menu item, find the saved PDB file and open it. The default view is 'Ball and Stick representation'. Rotate the molecule by holding down the left mouse button and moving the cursor.

12| To visualize the Cys residues and to locate the disulfide bonds, use the 'Select' command. Right-click in the molecule's window and expand the menus by moving the mouse cursor over to the menu items. Select 'Protein/by residue Name' and left-click on 'CYS'. Change the representation of the selected atoms by a right-click and expand the menus into 'Style/Atoms'. Click on '75% Van der Waals'. The atoms of the Cys residues will be seen as spheres and will be distinguished from the rest of the protein's atoms. The sulfur atoms are represented in yellow. If a disulfide bond exists between two sulfur atoms, the atoms will be seen in close proximity to each other (**Fig. 6**). Using this approach, visualization of the Cys residues allows the detection of (i) the presence of disulfide bonds, (ii) their spatial arrangement within the protein's 3D structure and (iii) their accessibility to solvent and reagents.

13| As some parts of the protein may be absent from the PDB file because they have not been experimentally determined, assess which segments are missing. Look at the protein's backbone representation and search for breaks in its structure, or examine the PDB file using a text editor; that is, determine whether some residue numbers are missing in the ATOM rows of the PDB file. If this is the case, the missing parts can be modeled using loop modeling or homology modeling. The modeling of missing sequences can be carried out using the homology modeling servers at the web addresses <http://www.expasy.org/swissmod/SWISS-MODEL.html> and <http://www.bmm.icnet.uk/people/paulb/3dj/form.html>. The loop modeling servers are found at <http://mordred.bioc.cam.ac.uk/~rapper/> and <http://protein.cribi.unipd.it/lobo/>. Although a detailed description of how to model the missing parts of the protein's structure is beyond the scope of this protocol, instructions and tutorials are provided at these websites.

14| Use additional tools to confirm the disulfide bonds detected in Step 12, if desired. These optional tools use the quantitation of the geometric parameters given in **Table 1**. Disulfide bonds can be detected using the Disulfide by Design software package. Start the program by clicking the menu item 'Start/Programs/DisulfidebyDesign'. Recall the previously saved PDB file by clicking the 'Load Structure' button and start the analysis by clicking the 'Run' button. This program searches for the relative proximity of any Cys residues, assesses their arrangement and checks whether pairs of residues have the geometry and the angles typical of a disulfide bond. The program will report several pairs of residues that may or may not be Cys. A disulfide bond is most likely to form if there are two paired Cys residues (**Fig. 7**).

15| Evaluate the accessibility of disulfide bonds to solvent using the MOLMOL software. Start the software by double-clicking the molmol desktop icon (if created in software setup) or by using Explorer to double-click the 'molmol.exe' file in the 'c:\xxxx\molmol' directory. Load the PDB file by selecting the menu item 'File ReadMol PDB' and finding the file at the saved location. To calculate the SASA, select the menu item 'Calc Surface'. A new dialog window will open. This is used to define the parameters for the calculation. The default values are acceptable. If a detailed report is required, the box called 'by atom' can be ticked. The software will then estimate the solvent accessible and the total surface area of the protein residues, followed by a percentage for the accessibility of each residue to solvent. Residues with a high percentage will have the greatest likelihood of being exposed to solvent. On the basis of our experience, we have classified a disulfide bond as 'buried' if the accessible surface of both of the Cys residues is less than 5%. If the SASA of one of the Cys residues in a disulfide bond is more than 10%, then it is reasonable to conclude that the disulfide bond is sufficiently accessible to solvent for it to be available for chemical

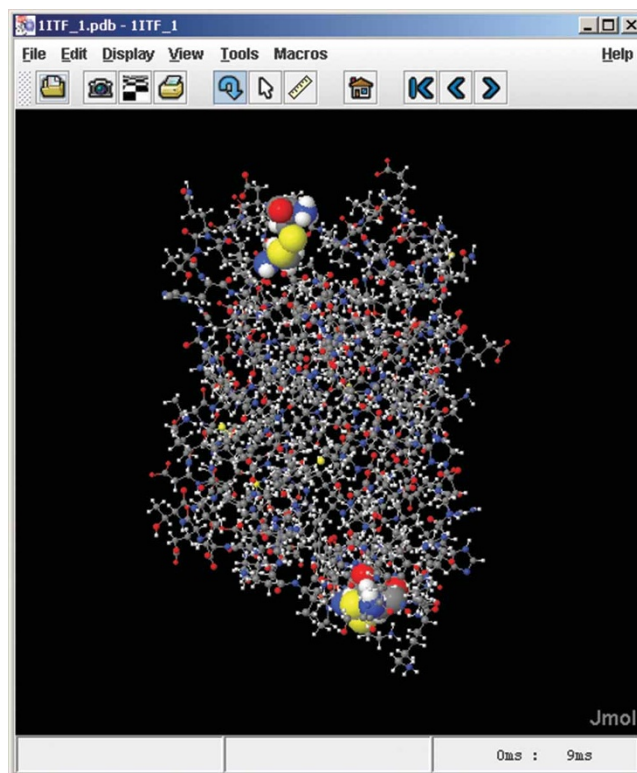


Figure 6 | Screen capture of Jmol software. This shows interferon α -2a with the Cys residues represented as Van der Waals surfaces.

reduction. The Cys residues that cannot be simply classified as being accessible to solvent will have to be examined using a combination of visualization and analytical tools.

▲ CRITICAL STEP Keep in mind that the PDB file is a static snapshot of a dynamic protein structure that is constantly moving and changing with time. This is especially true of the residues that are close to the surface. They have fewer stabilizing intra-protein interactions and they are more exposed to collisions with the solvent and other surrounding molecules. This means that some Cys residues that appear to be inaccessible might become intermittently exposed to the solvent and would therefore be available for chemical modification over a period of time. This dynamic situation can be investigated using molecular simulation of native proteins as described in part 2 of this protocol.

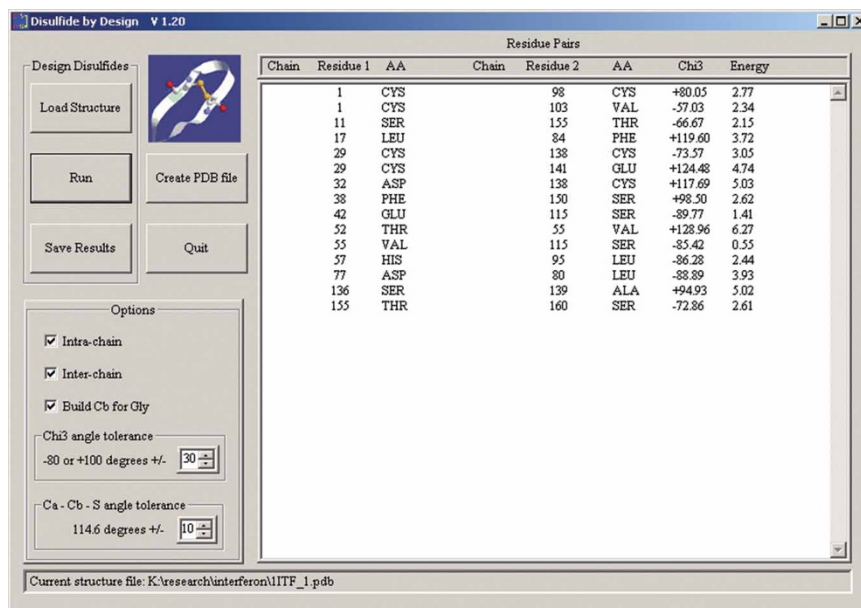


Figure 7 | The output of the analysis for the potential disulfide bonds that can form in the 3D model of interferon α -2a. This prediction was carried out using Disulfide by Design software. 'Chi3' refers to the torsional angle formed by C β -S γ -S γ -C β atoms. 'Energy' reports the interaction between the two Cys residues in kcal mol⁻¹. This makes it possible to compare the potential of several Cys residues to form a disulfide bond.

16 | After establishing that a protein has disulfide bonds, determine the change in its structure after the insertion of a 3-carbon bridge. A literature search for the ligand or receptor binding surface should aim to define the residues involved in receptor–ligand interactions. Using Jmol, it is also possible to examine the position of the modified disulfide bonds with respect to the active surface. Predicting the effects of the modifications on the protein's structure, with specific reference to its potential receptor or ligand binding interactions, is described in part 2 of this protocol.

Part 2: modeling the structural effects of modifying disulfide bond(s) ● **TIMING** Approximately 3 weeks—an experienced modeler could set up all of the calculations in 3 h

17 | For software validation if software other than MacroModel is utilized (optional), perform the following simulation on interferon α -2a in the first instance.

18 | Transfer the saved PDB file of the protein and resave it in a new location using the Linux file system.

19 | Open a Unix shell and change the directory's location to the file in which the protein's structure is saved. Start the Maestro package by typing 'maestro' in the Unix shell.

20 | Start a new project by selecting the 'Project/New' menu item and type a new name followed by clicking the 'Create' button. Use the 'Project/Import' menu item to insert the protein's structure into the workspace. A new window will appear in which you have to select the PDB format by clicking on the 'Format' dropdown menu; the default is Maestro. Select the filename with the protein's structure and click the 'Import' button.

21 | As atoms in the protein imported from the PDB format are usually gray in color, to distinguish different types of atoms, change the color of all the atoms using the menu item 'Display/Atom Coloring' and select 'Element'. Click on the 'All' button.

22 | Check if the hydrogen atoms are included in the PDB file, because most of the structures are derived from X-ray crystallographic studies and, as such, do not include hydrogen atoms. Therefore, the explicit hydrogen atoms have to be added to the structure (Step 25). In addition, the disulfide bonds are not created automatically. The simplest way to address these issues is to execute a sequence of scripts that have to be installed separately.

23 | Execute the script from the menu 'Scripts/Connect Disulfides'. It will connect all the S atoms from the Cys residues that are less than or equal to 3.2 Å apart.

24 | Confirm that the correct disulfide bonds have been created using the information found in the literature or from the results obtained using Disulfide by Design.

25| Prepare the protein for the simulation by using the script “Scripts/Protein Preparation Wizard” which offers several options. They can be used to prepare the protein for the simulation. The additional functions in the script will be available only if you have licenses for Epik and Glide. Click on the ‘Fix Structure’ button to assign the bond orders and to add the hydrogen atoms.

26| To evaluate whether there are any ligands in the structure or unusual residues, click the ‘Find Hets’ button. As these residues or ligands may be important to preserve the correct structure of the protein, an informed decision has to be made about whether they can be removed from the structure before carrying out the simulation. For example, three zinc ions are essential for the folding of the CBP/TAZ1 domain into a functional conformation (PDB entry 1U2N and references therein).

27| Click the ‘Run Protein Assignment’ button to optimize the states of the hydroxyl, Asp, His and Gln residues.

28| Use Epik to set the protonation states of the residues at the desired pH (optional).

29| If a Glide license is available, run a constrained refinement using ‘Run Impref Minimization’ to remove the initial steric clashes and to optimize the positions of the hydrogen atoms that were added in Step 22.

30| Save the resulting structure of the unmodified protein in the Maestro file format using ‘Project/Export’ and give it a new name. This structure can be used to run a simulation protocol and the results are used to determine whether the protocol is suitable for the protein. In the following steps, we describe the procedure that we used for the simulation of interferon α -2 by stochastic dynamics⁹. The same setup can be applied to all other simulations of the reduced and the modified protein.

31| Select the ‘Applications/Macromodel/Dynamics’ option from the menu. A new window will open that offers several different pages under tabs.

32| Select the tab named ‘Potential’. Using the dropdown menus, set up calculations using the OPLS_2005 force field and WATER solvent. This will, in turn, set up appropriate parameters for the dielectric constant charges (1.0) from force field, and extended cut-off values for Van der Waals (8.0), electrostatic (20.0) and H-bond interactions (4.0) Å.

33| If the simulation does not require constraints, select the ‘Constraints’ tab and press both ‘Reset All’ buttons. If simulation of the protein’s substructures is not required, select the ‘Substructure’ tab and clear the selection by pressing the ‘X’ button. The substructure feature should be considered if the size of the molecule exceeds the atom limit of Macromodel and the simulation fails. The atom limit is not clearly defined in the manual; for example, we could not carry out a simulation of asparaginase (a multimeric protein with more than 9,000 atoms) using all-atom force field. In such cases, we recommend that the PEG moiety and the region around the disulfide bond that requires modification be selected as a substructure.

34| On the ‘Mini’ page, select the ‘PRCG’ method and 2,500 as the maximum number of iterations. Set the convergence on ‘Gradient’ to a threshold of 0.05. This combination should provide a stable starting structure for the molecular simulation. Other values can be considered; however, a low number of iterations (fewer than 500) and a high convergence gradient (greater than 0.5) might not provide a stable starting structure for the dynamics—which can then fail. Otherwise, a small convergence gradient (less than 0.01) during the optimization step might not be achieved and the simulation step will not start.

35| On the ‘Monitor’ page, set the number of structures to be sampled to 200. Higher numbers are possible, but handling the structures becomes less practical. Another option is to set up additional monitors that will record selected distances, angles and hydrogen bonds.

36| On the ‘Dynamics’ page, use the dropdown menus to select the ‘Stochastic dynamics’ method for simulation and ‘Bonds to hydrogens’ SHAKE method. We have carried out simulations on 300 K with a timestep of 1 fs. The system was equilibrated for 10 ps with a simulation of 2,000 ps. It is important to note that the timestep of 1 fs in combination with a SHAKE constraint ‘Bonds to hydrogens’ is sufficiently small to prevent stretching bond lengths to unrealistic values. The 10 ps equilibrium is a period during which a system can equilibrate when it is heated from 0 to 300 K. If the protein is unstable (i.e., loses its tertiary structure) during this initial period of simulation, the equilibrium period can be set to longer values. For AMBER software, this equilibrium period can be up to 1,000 ps. We experimented with the duration of the simulation for reduced interferon α -2a and found that the RMSD value plateaus before 2,000 ps. For simulations with leptin, a shorter period of 1,600 ps was needed to achieve the plateau values of the RMSD. Therefore, our recommendation is to increase the duration of the simulation of the RMSD until a plateau value is reached.

37| Depending on the queue system used and personal preference, it is possible to proceed according to either option (A) or option (B) as described below. Schrödinger software provides support for the portable batch system, a load sharing facility and SGE queuing systems. The installation is described in the Maestro documentation. In either case, the results of the simulation are recorded in several different files. Of primary interest are the files ‘selected_name-out.mae’, which contains the 200 structures (i.e., snapshots of the simulation), and ‘selected_name.log’, which contains information on the progress of the minimization and simulation and records of total energies and temperatures every 10 ps during the simulation.

(A) Using the batch job submission

- (i) Press the 'Write' button and open a new window. The initial structure and batch (command) file will be written under the 'selected_name', which, in turn, can be submitted as a batch job as the command '\$SCHRODINGER/bmin selected_name'. The simulation cannot be monitored using Maestro. The Linux batch queue has to be interrogated to find out whether the job has finished.

(B) Using the queue system within Maestro

- (i) Press the 'Start' button. A new window will open and the simulation can be submitted under 'selected_name' to a queue that is defined in the dropdown menu for 'Host:'. The simulation can be followed through the 'Monitor' window (Applications/Monitor).

38| Analyze the simulation using freeware molecular graphics packages (option A) or Maestro (option B).

(A) Analysis using freeware packages

- (i) As freeware packages (VMD, VEGA ZZ) do not read the Maestro file format, convert the file into a PDB file format. Import the 'selected_name-out.mae' file with the 'Import all structures' and 'Replace Workspace' boxes checked. They are red in color. To save the files, use the 'Export' feature. From the dropdown menu, select the 'PDB' format. Check the box 'Selected entries', which is red in color. From the dropdown menu, select 'Export each entry as an individual file' with 'File name + automatic number' selected for the 'File names are' option. Give the file a name and press the 'Export' button. This will create 200 new PDB files that can be analyzed separately or concatenated into a single file for analysis.
- (ii) As the automatic numbers given during the export will result in the listing of the files being in non-numerical order, modify the filenames by adding '00' in front of each structure number for structures 1 to 9, and by adding '0' to the front of each structure number for structures 10–99. Packages such as VEGA, VMD and MOLMOL can be used for this analysis and for producing graphs of the RMSD.

(B) Analysis using Maestro

- (i) Use 'Project/Show Table' to open a new window in which structure names and their corresponding properties are shown in a table. Select a structure by clicking the diamond in the 'In' column in the corresponding row. The selected structure will be displayed in the graphics window. Selection of multiple structures can be achieved using the Shift or Ctrl keys on the keyboard and then clicking on the diamonds for the structures of interest. For example, the conformation of the protein at the start and at the end of the simulation can be compared by selecting the first and the last structure.
- (ii) Follow the dynamic behavior of the structures using the ePlayer options and by playing the trajectory forward or backward. If the first snapshot from the simulation is shown on the screen, the snapshots of the conformations taken at different time points during the simulation can be viewed by pressing 'Play' on the ePlayer. This will lead to a review of the changes in the protein's conformation with time.
- (iii) Follow the changes in protein structures and other properties that occur during simulation by calculating selected properties. In particular, RMSD and the distance between selected atoms are usually of greatest interest. For example, monitoring RMSD makes it possible to track the displacement of the atoms during the simulation as well as the deviation in an atom's position at specific times. To monitor RMSD values, select all of the structures in the trajectory. From the menu 'Tools/Superposition', select the atoms for which the RMSD will be calculated, for example a whole protein or only certain parts of the protein (e.g., active site) or a selected group of atoms (e.g., backbone). If the change in structure is of interest, the box 'Calculate in place (no transformation)' should not be selected because this will lead to the translation and rotational movements of the protein also being considered. For RMSD to be included in the 'Project Table', select the 'Create Property' box. The calculation of RMSD is carried out when the atoms are selected either by clicking the 'All' button or after selecting atoms using the 'Select' button. The numerical values of the RMSD for each conformation are reported in the bottom of the RMSD window.
- (iv) In a similar fashion, the distance between pairs of molecules for snapshots of the structure can be inserted into the 'Project Table'. From the menu, select 'Tools/Measurements', open the tab 'Distances' and tick the 'Create Property for Selected Entries' box. Pick a pair of atoms such as the two sulfur atoms for the disulfide bond. This will create a new column in the 'Project Table'.
- (v) The change of a particular property with time can be followed by creating a graph. Open the 'Project/Project Table' menu item. In the new window, select the 'Table/Plot' menu item. In the 'Plot XY' window, select the 'Plot/New Plot' menu item. This will open a 'New Plot' window. Write plot and series name in the appropriate boxes and then select the properties required—time for the x-axis and, as an example, RMSD for the y-axis. The drawing style on the graph can be further customized by selecting the desired options from the dropdown menus. The new graph will be created by clicking on the 'New' button, and the image can be exported in tiff format by selecting 'Plot/Save PlotXY Image'.

39| Also carry out simulation of the reduced protein. To reduce the disulfide bond, delete the bond between the sulfur atoms of the Cys residues of interest. The starting structure of the native protein is imported using the file created in Step 30. The disulfide bond that should be reduced can be located by hiding all of the atoms in the first instance (by selecting the 'Display/Undisplay Atoms' item on the Display menu) and clicking the 'All' button that corresponds to the 'Undisplay' row. To



visualize the required Cys residue, click the 'Select' button that corresponds to the 'Also Display' row. A new window will appear. Under 'Residue' tab, choose the Residue number selection criteria. In the Residue number, type the number of Cys residues that correspond to the disulfide bond requiring reduction. Click the 'Add' button and 'OK'. The selected atoms will be displayed with the disulfide bond(s) in yellow. Click on the button 'X' on the left-hand side of the graphics window and hold the mouse button down until the menu appears. Drag the mouse pointer to the bonds and release the mouse button. The click of the mouse button will change the function to deleting bonds. Click the bond between the two yellow sulfur atoms and it will disappear. Double-click on the 'H' button on the graphics menu to add hydrogen atoms to the whole molecule. This will also add hydrogen atoms to the sulfur atoms. The modified molecule will correspond to the protein with the reduced disulfide bond. Repeat this for all the disulfide bonds that have to be reduced. Save the reduced structure as a new file in Maestro format ('Project/Export').

40| Carry out the simulation protocol defined in Steps 31–37 and analyze the trajectory as described in Step 38.

41| Using builder, delete hydrogen atoms on the sulfur atoms of the reduced disulfide bond. Use the freehand drawing tool 'Draw Structure' on the build window to connect the sulfur atoms with a 3-carbon bridge. Click on the first sulfur and then click another three times to draw three connected carbons. Finally, click on the second sulfur atom. Save the structure under a new name.

▲ CRITICAL STEP Since the view on the screen is a 2D representation of a 3D conformation, care is required to ensure that the chain connects the correct atoms. Mistakes will lead to defective topologies. View all atoms by selecting the menu item 'Display/Undisplay Atoms' item on the Display menu and in the new window click the 'All' button in the section 'Also Display'. By holding down the middle button (or wheel) and the right button on the mouse, zoom the view of the modified disulfide bond and make sure that the new carbon chain is connected to the correct sulfur atoms. It is also important to make sure that it is along the same line as the original disulfide bond. In addition, it should not clash with other atoms in the protein.

42| Carry out the simulation protocol defined in Steps 31–37 and analyze the trajectory as described in Step 38.

43| If PEGylation is desired, start by using the molecular builder to model the PEG molecule on its own. PEG is already folded when it undergoes reaction with a protein. Start with an empty workspace. Select 'Edit/Build' and from the 'Fragments' dropdown menu select 'Organic'. Pick 'Place' box and click on the methyl fragment structure. This will add a methane molecule, and it is the starting point for building a PEG chain with a terminal methyl group. Pick the 'Grow' box and click on the 'hydroxyl' fragment to add oxygen atoms to the chain. This is followed by two clicks on the 'methyl' fragment. This will complete the addition of a glycol monomer. Repeat the process by adding one hydroxyl and two methyl fragments until the PEG has grown to the desired size. We have used PEGs of 1–10 kDa. Finish with the final hydroxyl group and one more methyl group. This will create an extended chain that has to be folded before it can interact with the protein.

44| Generate the conformation of the PEG moiety to be connected to the protein. Having tested several different simulation and conformational search protocols, we have found that a PEG moiety exhibits similar structural features irrespective of whether the folding studies have been carried out using explicit or implicit solvent (water) representation. The most efficient way of generating a PEG conformation is to use a modified simulation protocol for the protein as defined in Steps 31–37. Use the molecular dynamics simulation instead of stochastic dynamics. After the simulation, use one of the conformations obtained during the last few picoseconds of the simulation to model the PEGylating moiety. A linker has to replace the terminal methoxy group. Choose the end that is not buried inside the PEG and delete the terminal methoxy group and the oxygen atom. Define a grow bond by selecting the last two carbon atoms. Add an amine fragment followed by a carbonyl group, phenyl ring and then one more carbonyl group. This will complete the PEG with an amine linkage. Save the structure using the 'Export' function.

45| To model a PEGylated protein, use the structure of the protein with the modified disulfide bond into which a 3-carbon bridge has been inserted as described in Step 42. Either use the initial modified structure of the protein or import one of the conformations obtained during the simulation (Step 43). Do not use an average structure from the set of snapshot conformations because it may not have realistic coordinates.

46| Import the PEG linkage (without replacing the protein structure) into the workspace. Click the 'Local transformation' button on the graphical menu and click the PEG. This will allow the PEG to be rotated and translated without moving the protein. Use rotation and translation to dock the PEG onto the protein in such a way that the linkage on the PEG is in close proximity to the 3-carbon methylene chain. Ensure that steric collision between the PEG and the protein does not occur. This will allow connection of the middle carbon atom of the bridge to the carbon atom on the terminal carbonyl group on the linkage. Double-clicking on the 'H' button will add the required hydrogen atoms and complete the structure. Save the PEGylated protein under a new name and then submit it to simulation using the protocol as defined in Steps 31–37.

47| After the simulation, analyze the trajectory using the tools described in Step 38. Also, use these tools to compare the structure of the native protein, the protein with reduced disulfide bonds and the PEGylated protein.

48 | Examine the effect of PEGylation on the structure of the active surface by calculating the RMSD [Step 38B(iii)] for the atoms that make up the active surface of the native protein and the modified protein.

? TROUBLESHOOTING

● **TIMING**

Steps 1–3: 0.5 h

Steps 4–10: 0.5 h

Steps 11–13: 0.5 h unless additional modeling is required

Step 14: 0.5 h

Step 15: 0.5 h

Step 16: 0.5 h

Note: The timing for the rest of the procedure will depend on the hardware, the software setup and previous experience. The times given below are for a Suse Linux 9.3 system running Maestro version 6.5 and Macromodel version 9.11 on an Opteron 2.2 GHz processor. These timings will change if the parameters of the simulations are changed from the values given in this protocol.

Steps 17–19: 0.5 h

Steps 20–30: 1 h

Steps 31–37: 5 d for 146 aa residues; that is, the time taken will depend on the size of the protein

Step 38: analysis will depend on experience and the detail required

Steps 39–41: as Steps 31–38

Step 42: 0.5 h

Step 43: as Steps 31–38

Steps 44–45: time required for this step will depend upon the size of the PEG built; allow 3 h for each 1 kDa of PEG

Steps 46–47: 7 d for a 146-aa protein and a 10-kDa PEG

Steps 48: analysis will depend upon experience and detail of information required

? TROUBLESHOOTING

Troubleshooting advice can be found in **Table 3**.

TABLE 3 | Troubleshooting table.

Problem	Possible reason	Solution
No structural information is available for the protein	Protein is not crystallized or too large for NMR, etc.	Consider homology or <i>ab initio</i> protein modeling using the known sequence. This will require specialist molecular modeling expertise
Some parts of the structure of the protein are missing in the PDB file	Some segments of the protein chain are flexible. This prevents their experimental determination	Consider adding the missing parts by loop modeling or homology modeling
<i>Macromodel specific</i>		
Molecular simulation fails to run	Size of the protein exceeds the program's atom limit	Run dynamics on relevant substructures Use AMBER-united atom force field
	The starting structure for dynamics simulation is not stable	Increase the number of iterations in the Mini tab of the Dynamics setup Use a shorter time step When setting the dynamics, select the 'Bonds to all atoms' option in the Shake method
Native structure unfolds in the early steps of the molecular simulations	System not equilibrated	Increase equilibrium period

ANTICIPATED RESULTS

Part 1: structural analysis

After completing part 1 of the procedure, the user will have information about the existing knowledge base for the structure of the protein being studied, the number of Cys residues and their mutual interaction as disulfides. In many cases, it will be possible to determine and visualize the location of the disulfide bonds in the protein.

As an example, the sequence of interferon α -2 is shown in **Figure 3**. Six Cys residues are present. Two Cys are in the precursor leader sequence. From the references cited, it is known that biologically active interferon α -2 has two disulfides: Cys1–Cys98



and Cys29–Cys138, with the numbering excluding the signal sequence shown in **Figure 3**. In the case of proteins for which information does not exist about which Cys residues are involved in disulfide formation, computational strategies can be used to predict (with varying degrees of accuracy) whether each Cys could pair to form a disulfide bond^{25,26}. It is a good practice to recheck that the entire sequence of the desired protein is reported in the PDB file because some sections of the protein may be missing. For example, **Figure 4** shows the available X-ray structure of leptin (PDB entry 1AX8). It does not include the flexible parts of the protein.

TABLE 4 | Root mean square deviation (RMSD) values for native and modified proteins and their correlation with the observed folding state of the modified protein.

RMSD (Å)	Folding state of the modified protein
< 4.5	Within the conformational flexibility of the native protein
5–10	Partial unfolding of the modified protein is likely to occur
> 10	Complete unfolding of the modified protein will occur

Part 2: the modeling of the reduction and the rebridging of disulfide bonds

The simulation protocol should provide the user with sufficient information about the structure and dynamic behavior of the native protein, the reduced protein and the modified protein. With this knowledge, it becomes possible to make an informed decision about the suitability of the protein for site-specific modification by the insertion of a 3-carbon bridge across the two sulfur atoms of a reduced disulfide bond. Such a protein has the potential to be a target for site-specific PEGylation using our protocol¹¹.

Our modeling experiments have shown that reducing disulfides in different proteins can have a variety of effects on their structure⁹. Other published studies have also described the effects of reducing disulfides on a protein's structure. For example, a 2-ns molecular dynamics simulation of azurin indicated that its overall structure was not affected by removing the disulfide bond²⁷. This was in contrast to the 5-ns molecular dynamics simulations of a 47-residue peptide, thionin, which has four native disulfides. Its tertiary structure was significantly affected by the reduction of just one disulfide²⁸. Our experimental studies have shown that the structures of interferon α -2 and an anti-CD4 Ab do not change significantly after the reduction of an accessible disulfide bond¹⁰. We were able to run nanosecond simulations to confirm that the structures of these proteins were retained after the reduction of a disulfide⁹. When calculating RMSD values, consider the parts of the protein that have well-defined structural features. In the first instance, exclude flexible loops at the C- or N-terminus. The RMSD values of the native conformation and the modified conformation of interferon α -2a during our simulations did not exceed 4.5 Å. As such, they were within the limits of the RMSD values calculated using the various experimental NMR models of interferon α -2a. It is important to realize that movement of the protein subunits (or domains) that are connected by flexible loops can result in large changes in the RMSD without significantly affecting the protein's secondary structure. Although it is difficult to define with precision the changes in RMSD that will correlate with a disruption of a protein's tertiary structure, we have attempted to do so for our observations in **Table 4**.

When studying a protein for which experimental results are not available, we recommend that the simulations are run for longer periods of times than the literature-based cases mentioned above. The running times for proteins for which little information is available can range from tens of nanoseconds to several milliseconds²⁹. Note that these longer simulations will require access to substantial computational power.

We have already described how parts of the 3D structure of leptin were missing from the PDB (entry 1AX8; **Fig. 4**). After modeling these missing parts and carrying out the simulation, we found that the protein's structure was not significantly affected by the reduction of a disulfide bond (i.e., the RMSD of the reduced protein was less than 3.5 Å when compared with the native conformation of leptin), but there was increased flexibility of the C-terminus of the protein, in which one of the Cys residues was present. The distance between the two sulfur atoms was 4 Å greater than the distance between the sulfur atoms in the disulphide bond. Experimental data confirmed that reduction of the accessible disulfide bond did not lead to denaturation of the protein (S. Balan and S. Brocchini, unpublished results).

ACKNOWLEDGMENTS This work was supported by the Biotechnology and Biological Sciences Research Council (BBSRC) UK (BB/D003636/1) and the Wellcome Trust (068309).

COMPETING INTERESTS STATEMENT The authors declare no competing financial interests.

Published online at <http://www.natureprotocols.com>
Reprints and permissions information is available online at <http://npg.nature.com/reprintsandpermissions>

1. Betz, S.F. Disulfide bonds and the stability of globular proteins. *Protein Sci.* **2**, 1551–1558 (1993).
2. Petersen, M.T., Jonson, P.H. & Petersen, S.B. Amino acid neighbours and detailed conformational analysis of cysteines in proteins. *Protein Eng.* **12**, 535–548 (1999).

3. Bhattacharyya, R., Pal, D. & Chakrabarti, P. Disulfide bonds, their stereospecific environment and conservation in protein structures. *Protein Eng. Des. Sel.* **17**, 795–808 (2004).
4. Rajarathnam, K., Sykes, B.D., Dewald, B., Baggiolini, M. & Clark-Lewis, I. Disulfide bridges in interleukin-8 probed using non-natural disulfide analogues: dissociation of roles in structure from function. *Biochemistry* **38**, 7653–7658 (1999).
5. Bewley, T.A., Brovetto-Cruz, J. & Li, C.H. Human pituitary growth hormone. Physicochemical investigations of the native and reduced-alkylated protein. *Biochemistry* **8**, 4701–4708 (1969).
6. Wedemeyer, W.J., Welker, E., Narayan, M. & Scheraga, H.A. Disulfide bonds and protein folding. *Biochemistry* **39**, 4207–4216 (2000).
7. Morehead, H., Johnston, P.D. & Wetzel, R. Roles of the 29–138 disulfide bond of subtype A of human alpha interferon in its antiviral activity and conformational stability. *Biochemistry* **23**, 2500–2507 (1984).



8. Todokoro, K., Saito, T., Obata, M., Yamazaki, S. & Tamaura, Y. Studies on conformation and antigenicity of reduced S-methylated asparaginase in comparison with asparaginase. *FEBS Lett.* **60**, 259–262 (1975).
9. Godwin, A. *et al.* Molecular dynamics simulations of proteins with chemically modified disulfide bonds. *Theor. Chem. Acc.* **117**, 259–265 (2007).
10. Shaunak, S. *et al.* Site-specific PEGylation of native disulfide bonds in therapeutic proteins. *Nat. Chem. Biol.* **2**, 312–313 (2006).
11. Brocchini, S. *et al.* PEGylation of native disulfide bonds in proteins. *Nat. Protoc.* **1**, 2241–2252 (2006).
12. Thornton, J.M. Disulphide bridges in globular proteins. *J. Mol. Biol.* **151**, 261–287 (1981).
13. Balan, S. *et al.* Site-specific PEGylation of protein disulfide bonds using a three-carbon bridge. *Bioconjug. Chem.* **18**, 61–76 (2007).
14. Liberatore, F.A. *et al.* Site directed modification and crosslinking of a monoclonal antibody with equilibrium transfer alkylating crosslink reagents. *Bioconjug. Chem.* **1**, 36–50 (1990).
15. del Rosario, R.B., Wahl, R.L., Brocchini, S.J., Lawton, R.G. & Smith, R.H. Sulfhydryl site-specific cross-linking and labeling of monoclonal antibodies by a fluorescent equilibrium transfer alkylation cross-link reagent. *Bioconjug. Chem.* **1**, 51–59 (1990).
16. Wilbur, D.S., Stray, J.E., Hamlin, D.K., Curtis, D.K. & Vessella, R.L. Monoclonal antibody Fab' fragment cross-linking using equilibrium transfer alkylation reagents. A strategy for site-specific conjugation of diagnostic and therapeutic agents with F(ab')₂ fragments. *Bioconjug. Chem.* **5**, 220–235 (1994).
17. Floudas, C.A., Fung, H.K., McAllister, S.R., Mönnigmann, M. & Rajgaria, R. Advances in protein structure prediction and *de novo* protein design: a review. *Chem. Eng. Sci.* **61**, 966–988 (2006).
18. Dombkowski, A.A. Disulfide by Design: a computational method for the rational design of disulfide bonds in proteins. *Bioinformatics.* **19**, 1852–1853 (2003).
19. Chang, S.G., Choi, K.D., Jang, S.H. & Shin, H.C. Role of disulfide bonds in the structure and activity of human insulin. *Mol. Cells* **16**, 323–330 (2003).
20. Koradi, R., Billeter, M. & Wuthrich, K. MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graph.* **14**, 51–55 (1996).
21. Zloh, M., Balan, S., Shaunak, S. & Brocchini, S. The effect of hydrogen bonding interactions on the reactivity of a novel disulfide-specific PEGylation reagent. *8th International Conference on Fundamental and Applied Aspects of Physical Chemistry* (Belgrade, Serbia, 2006).
22. Zloh, M., Balan, S., Shaunak, S. & Brocchini, S. Modeling study of disulfide bridged PEGylated proteins. *6th European Conference on Computational Chemistry* (Slovakia, 2006).
23. van Gunsteren, W.F. & Berendsen, H.J.C. A leap-frog algorithm for stochastic dynamics. *Mol. Simul.* **1**, 173–185 (1988).
24. Shen My, M.Y. & Freed, K.F. Long time dynamics of Met-enkephalin: comparison of explicit and implicit solvent models. *Biophys. J.* **82**, 1791–1808 (2002).
25. Martelli, P.L., Fariselli, P., Malaguti, L. & Casadio, R. Prediction of the disulfide-bonding state of cysteines in proteins at 88% accuracy. *Protein Sci.* **11**, 2735–2739 (2002).
26. Ferre, F. & Clote, P. DiANNA: a web server for disulfide connectivity prediction. *Nucleic Acids Res.* **33**, W230–W232 (2005).
27. Rizzuti, B., Sportelli, L. & Guzzi, R. Evidence of reduced flexibility in disulfide bridge-depleted azurin: a molecular dynamics simulation study. *Biophys. Chem.* **94**, 107–120 (2001).
28. Vila-Perello, M. & Andreu, D. Characterization and structural role of disulfide bonds in a highly knotted thionin from *Pyrularia pubera*. *Biopolymers* **80**, 697–707 (2005).
29. Snow, C.D., Nguyen, H., Pande, V.S. & Gruebele, M. Absolute comparison of simulated and experimental protein-folding dynamics. *Nature* **420**, 102–106 (2002).