In Silico Transfer of Ligand Binding Function between Structurally Analogous Proteins

Ali Awan

Introduction

One of the major goals of biotechnology is the deliberate modification of proteins, or protein engineering, for scientific, industrial and medicinal purposes. The earliest applications have been in scientific studies, in which protein engineering has helped further our knowledge of proteins and their chemical makeup. More recently, deliberate protein modification has been used to alter protein performance for industrial purposes. It has been possible to change the rates, temperatures and pHs at which enzymes operate, to increase efficiency and applicability. Further, it has become possible to modify not only protein performance, but even protein function. Some specific examples include producing enzymes to degrade industrial plastic waste, exploiting the tight binding of antibodies to other molecules to make them catalytic, and engineering proteins to bind and inactivate virii.

The main methods of protein engineering employ organic and physical chemistry as well as genetic engineering. Chemical protein engineering can involve reactions of the constituent amino acid side chains, active site directed reactions, protein backbone cleavage and protein synthesis using proteases[1] . On a genetic engineering level, the most commonly applied protein engineering makes use of site directed mutagenesis. This entails performing PCR with one perfectly complementary primer and one primer that contains the desired mutations, leading to the production of mutated single DNA strands[2].

With the advent of increasingly sophisticated computational biology methods, it has become feasible to predict protein folding, model protein ligand docking, and identify protein motifs with increasing accuracy. Though these methods are still limited[3], and by no means infallible compared to their laboratory equivalents, they can be applied for '*in silico* protein engineering'. Specifically, one can use currently available computational methods to predict the stability and functionality (in terms of ligand binding) of proteins arising from modified amino acid sequences. This would either provide a starting point for further laboratory studies, or for the *in vitro* creation of the protein determined *in silico*.

There are two main advantages of this type of preliminary *in silico* protein engineering. First, it is likely to be much faster and less painstaking to test the effects of altered amino acid sequences on protein structure and function using computational methods rather than laboratory procedures. Second, there are no limits to the types of amino acid sequence one can explore *in silico*. Amino acid sequences in natural proteins result from natural selection, and thus are likely to be situated on 'evolutionary peaks'. It is unlikely that other amino acid sequences which are equally functional but situated on different evolutionary peaks are present in nature, due to the existence of 'evolutionary troughs' between the peaks. *In silico* methods can be used to design proteins based on stability and functionality, even if they do not occur anywhere in nature.

One particular protein engineering application of recent interest has been the transfer of a specific function between two related proteins. For example, a bicarbonate binding function has been successfully transferred in vitro from crocodile hemoglobin into a human-crocodile hybrid hemoglobin, with a minimal number of amino acid changes[4]. In this paper, I propose a general method for the *in silico* transfer of ligand binding function between structurally analogous proteins, and then examine its application to the specific example mentioned above.

General Procedures:

The protein with the desired ability to bind the ligand (I will term this 'functionality' for convenience), shall be referred to as A and the other protein as B. The ligand will be called L. To be able to evaluate the success of particular hybrids in retaining A's ability to bind L it is necessary to define a threshold percentage of acceptable functionality (T). T can be specified on a case by case basis, depending on the intended goals of the *in silico* function transfer.

Depending on how well characterised the protein type and ligand binding site are, there are several possible places to start. I will represent two different possible procedures in

pseudo-flow chart form, with explanations of each step. The methods for analysing hybrid stability and binding affinity to L are outlined below.

To ascertain the stability of a hybrid protein, the amino acid sequence is entered into an energy minimisation and protein folding program for engineered sequences, such as genome@home. This scores the stability of the protein, allowing the ranking of different hybrids in terms of stability. To evaluate the affinity of an engineered protein for the ligand, a docking program is used. Examples of docking programs are Dock, AutoDock and FlexX. These programs take into account ligand flexibility, protein surface flexibility, efficient sampling of the ligand conformational space, and sufficiently accurate energy functions to evaluate the best protein-ligand association[5] . The solutions can be ranked according to interaction energy.

Currently, the computational complexity of predicting ligand protein docking[6] means that the user has to specify the binding site to be used in the calculations. Thus one cannot simply enter the whole protein into the docking program and expect it to find the best place for the ligand to bind, out of all possible conformations in all protein locations. However, in the future, this might become a feasible goal of docking programs. With this in mind, I have proposed two different approaches to the transfer of function between analogous proteins *in silico*: one that assumes the feasibility of entering the whole protein into the docking program to find the best possible protein-ligand association and one that assumes the current docking program limitation (that the binding site must be user specified). These have been termed procedure one and procedure two respectively. Procedure two more closely mimics the in vitro procedure followed by Nagai et al.[7] for the transfer of bicarbonate binding function from crocodile to human hemoglobin.

Procedure One:

1. ***Is a motif for the binding site of L to A already characterised?***

To find out, search a database of motifs for ligand binding sites, such as ligbase, for the binding of L to proteins of type A. If there is no functional motif for the binding of L to

A, go to 2. If, on the other hand, a functional motif for the binding of L to A is already known, search for it in A using any string search method. Once found, this region in A should be the subject of the procedure in 5 to 7.

2. ***Is the general structural location (and hence the locations in the amino acid sequence) of the binding site of L to A known?***
If not, go to 3. Otherwise, carry out the procedure in 4 to 7 using the amino acid sequence(s) that form the region of the binding site in A.

3. ***Do A and B have structural subunits?***
If not, go to 4. Otherwise, start with the structure of A and replace each subunit one at a time with the corresponding subunit from B. For each hybrid thus produced, ascertain protein stability and ligand affinity, to check if the functionality is above the threshold functionality.

This way, one determines the subunits in A that are necessary for the ligand binding. It is much easier to replace A's subunits with B's to examine loss of function than trying to determine which subunits are necessary by attempting to reconstruct the function in B by sequentially replacing B's subunits with those of A.

It is possible that multiple subunits may be necessary for the ligand binding because of the effects of one subunit one the other. For example chemical properties of amino acids in one subunit could influence amino acids in the actual binding site in another subunit to arrange themselves in a way conducive to ligand binding. Once the necessary subunits have been determined, go to 4.

4. Once functionality has been localised to specific regions of A (call these regions X), one must determine which amino acids in these regions should and should not be mutated when determining those amino acids necessary for functionality. This saves computational time by obviating the need to mutate every single amino acid position.

To do this, use an application of the Smith Waterman algorithm, such as BESTFIT, to find the local alignments between regions X in A and the corresponding regions in B. In the locally aligned regions, search for structural motifs, using a program with protein motif finding capabilities, such as eMotif. If any are found, align each structural motif locally, and examine each amino acid position in the two sequences.

Those positions in the motifs that have the same exact amino acid will not be mutated. All other amino acids (call them M), both those that are in the structural motifs but different and those in X but not in any structural motifs (even if they are identical in A and B), are subjected to the procedure in 5. The amino acids in M in the example below are coloured blue.

Hypothetical example:

Local alignment of part of X in A and B:

A: APQCTHBEDILVKYCAS

B: APQRTHBEDVRVKYCEG

Structural motif found: HBED*[RL]VKY

Align structural motifs:

A: HBEDILVKY

B: HBEDVRVKY

5. Using the amino acid sequence of A, change each amino acid in M, one at a time, trying all 20 amino acids for that position. For each hybrid thus produced, evaluate the functionality using a docking program. Depending on how many substitutions there are in a given position for which functionality is not compromised (i.e. functionality does not go below the threshold level) each position can be determined as having high, medium or low 'information content' 5^{th} .

For example, if a certain amino acid originally present in M cannot be changed to any other amino acid without functionality being compromised, it is said to have high information content. On the other hand, if any amino acid can be substituted without

functionality being compromised, that position has low information content. If only substitutions of certain amino acids in that position (for example substitutions that are likely to occur in nature) do not affect functionality, the information content of that position is medium. Once all possible substitutions have been made for each position in M, and the information content of each position has been determined, go to 6.

Hypothetical example:

APQCTHBEDILVKYCAS

Mutate each blue amino acid position, using every possible amino acid for that position. For example, mutate the first position to G:

GPQCTHBEDILVKYCAS

Test this hybrid for functionality, as outlined above. Repeat until all amino acids have been tried in the first position, then assign information content level to the first position. Repeat for each position in M.

6. All the positions that were determined in 5 to have high information content in M are the ones that are essential to functionality (call them E). Transplant E from A into B (i.e. replace the corresponding amino acids in B with those from A), to create an initial hybrid sequence. Fold this hybrid and perform an energy minimisation on it using a program such as genome@home, and test it for stability and functionality. Go to 7.

Hypothetical example:

A: APQCTHBEDILVKYCAS

hlmhh mh hhl

(h means high information content, m means medium and l means low)

Hybrid: APQCTHBEDILVKYCAG

(red amino acids came from A, purple came from B, and black came from B but in positions that were not tested for information content)

Test for stability and functionality

7. If the hybrid produced in 6 is unstable, or its functionality is below the threshold level, two things can be done. First, those amino acids in the positions in M that had medium information content should also be transplanted from A to the hybrid. Second, if the

hybrid is still not stable or does not have threshold functionality, the subunits in the hybrid that are unmodified from those found in B (i.e. those that were not involved in steps 4 and 5) should be replaced by the corresponding ones in A, one at a time, and stability should be re-evaluated.

Procedure Two:

1. ***Is a motif for the binding site of L to A already characterised?***
To find out, search a database of motifs for ligand binding sites, such as ligbase, for the binding of L to proteins of type A. If there is no functional motif for the binding of L to A, go to 2. If, on the other hand, a functional motif for the binding of L to A is already known, search for it in A using any string search method. Once found, this region in A should be the subject of the procedure in 5 to 7.

2. ***Is the general structural location (and hence the locations in the amino acid sequence) of the binding site of L to A known?***
If not, go to 3. Otherwise, carry out the procedure in 4 to 7 using the amino acid sequence(s) that code the region of the binding site in A.

3. Use a program such as MOE Site Finder[8] to determine the potential binding sites of L to A. These potential binding sites will be used both as the sites to be mutated in A to determine which amino acid positions are necessary for functionality, and also as the binding sites that will be used in the protein ligand docking program calculations. If there is a significant difference between the scores of the top few solutions (even the top one) and the rest of them, use each of these top solutions for steps 4 to 6. Otherwise, try every solution given by the site-finding program in steps 4 to 6. Each potential binding site will be considered one at a time.

Steps 4 through 7 are identical to those in Procedure One.

Hemoglobin Example:

Since procedure one is not yet feasible with current docking programs, I will outline the hemoglobin example using procedure two. In this case, crocodile hemoglobin is A, human hemoglobin is B and bicarbonate is L, to use the terminology adopted in the general procedure.

However, the functionality is not simply how well bicarbonate binds to a hybrid, but is extended to include the molecule's affinity for oxygen with varying oxygen partial pressures (i.e. oxygen binding curves) in the presence and absence of bicarbonate. Since the oxygen binding site on hemoglobin is known, even though determining the oxygen binding curves is more complex than just determining the bicarbonate affinity, it can be evaluated using current docking programs. This is done by allowing multiple ligands (oxygen and bicarbonate) and varying the partial pressure of oxygen in the presence and absence of a set amount of bicarbonate.

Threshold functionality in this case is defined as a statistically significant difference between the oxygen binding curves of the hybrid in the presence and absence of bicarbonate, comparable to the difference in oxygen binding curves for crocodile hemoglobin under similar conditions.

1. Start by searching ligbase for a structural motif for the binding of bicarbonate to hemoglobin. Since this search returns no results[9] , we move to step 2.
2. The general binding site of bicarbonate to crocodile hemoglobin has already been determined to be somewhere in the alpha 1 – beta 2 subunit interface[10] .
3. Carry out steps 4 through 7 for the sequences that lie in the alpha 1 beta 2 subunit interface.
4. Compare the resulting best hybrid(s) for functionality with the empirically determined hybrid made by Nagai et al., both by *in silico* methods (using docking programs as outlined to test functionality in terms of oxygen binding curves) and *in vitro* methods (synthesize the *in silico* determined hybrids and then determine

their oxygen binding curves in the presence and absence of bicarbonate with *in vitro* techniques). This will serve as a test of the efficacy of the *in silico* transfer of ligand binding function between structurally analogous proteins.

Conclusions

At the end of the day, the general procedure for in silico transfer of ligand binding function between structurally analogous proteins is only as feasible as the computational methods are accurate. That is to say, the hybrids resulting from the in silico engineering will only resemble their empirically determined counterparts as far as the protein folding, energy minimisation and protein ligand docking algorithms manage to accurately predict the outcomes of the analogous in vitro processes.

Currently, the level of accuracy leaves a lot to be desired, especially when it comes to predicting structure based on sequence and predicting the best conformation for protein ligand docking[11]. Studies have shown that the accuracy of docking programs even when it came to simple ligands with known binding sites ranged from 30-70% [12] . Further, there are many factors that are not yet taken into account with current *in silico* methods, such as in the energy functions used by such programs: "presumably, the experimental structure … is defined by a fine balance of energy terms which is still beyond the accuracy of the available energy approximations" [13].

However, *in silico* methods for predicting protein folding and protein ligand docking are the subject of intensive current research[14] , and it is hoped that eventually the wide scale application of various types of protein engineering, not just the type examined in this paper, will be more than just theoretically feasible.

References:

1. Protein Engineering Introduction & Lecture Summaries http://www.bi.umist.ac.uk/users/mjfajdg/2PAB/default.asp
2. Protein Engineering Introduction & Lecture Summaries http://www.bi.umist.ac.uk/users/mjfajdg/2PAB/default.asp
3. Totrov, M., Abaygan, R., Flexible Protein-Ligand Docking by Global Energy Optimization in Internal Coordinates. PROTEINS Suppl 1;218-219 (1997)
4. Nagai, K., Tame, J., Komiyama, N., A Hemoglobin-Based Blood Substitute: Transplanting a Novel Allosteric Effect of Crocodile Hb. Biol. Chem 1996 Sep; 377(9):543-8
5. Totrov, M., Abaygan, R., Flexible Protein-Ligand Docking by Global Energy Optimization in Internal Coordinates. PROTEINS Suppl 1;215 (1997)
6. Totrov, M., Abaygan, R., Flexible Protein-Ligand Docking by Global Energy Optimization in Internal Coordinates. PROTEINS Suppl 1;215 (1997)
7. Nagai, K., Tame, J., Miyazaki, G., Komlyama, N. H., Transplanting aunique allosteric effect from crocodile into human haemoglobin. Nature 1995 Jan 19;673(6511):244-6
8. Labute, P., Santavy, M., Locating Binding Sites in Protein Structures. http://www.chemcomp.com/feature/sitefind.htm
9. Based on a search I carried out
10. Nagai, K., Tame, J., Miyazaki, G., Komlyama, N. H., Transplanting aunique allosteric effect from crocodile into human haemoglobin. Nature 1995 Jan 19;673(6511):244
11. Totrov, M., Abaygan, R., Flexible Protein-Ligand Docking by Global Energy Optimization in Internal Coordinates. PROTEINS Suppl 1;218-219 (1997)
12. Totrov, M., Abaygan, R., Flexible Protein-Ligand Docking by Global Energy Optimization in Internal Coordinates. PROTEINS Suppl 1;215 (1997)
13. Totrov, M., Abaygan, R., Flexible Protein-Ligand Docking by Global Energy Optimization in Internal Coordinates. PROTEINS Suppl 1;219 (1997)
14. Havranek, J., Harbury Research Group. http://cmgm.stanford.edu/biochem/harbury/people.html