

Using a Genetic Algorithm to Generate Decoy Sets for Protein Structure Prediction

Michael Sykes
D109 Fairchild Building
Stanford, California 94305
sykes@stanford.edu 5-0754

Abstract: In protein structure prediction, one method for determining a protein's three-dimensional structure is by selecting the most favorable conformation from a Decoy Set of possible structures. Here, a Genetic Algorithm using method-specific mutational operators is used to generate Decoy Sets for protein structure prediction. The method is tested on proteins whose detailed three-dimensional structure is already known. The best structure from each of these Decoy Sets is compared to the native structure, and the Decoy Sets are evaluated using different scoring potentials. The Genetic Algorithm demonstrates that it can predict some reasonable conformations, and that certain scoring potentials are able to correctly choose the best structure.

Note: This paper assumes some knowledge of protein structure and biochemistry. The author recommends Voet and Voet's *Biochemistry 2nd Edition* for those who would like to review the subject.

Introduction

The field of structural biology was for many years dominated by the actual determination of structure via experiment, usually X-ray crystallography or NMR spectroscopy. However in recent years, with the advent of the Human Genome Project and the mass sequencing of the genomes of other organisms, the face of structural biology has changed. We are now in an era where the number of three-dimensional structures of macromolecules is disproportionately small when compared to the number of sequences of macromolecules that are known. This is true of DNA, RNA and proteins, but is most apparent in the last case, and this condition continues to worsen. With no immediate hope of structure determination catching up to sequence determination, the field of structure prediction has flourished. All types of structures are predicted, but the protein structure predictors now dominate the landscape of computational structural biology.

The field has become so popular and dynamic that a protein structure prediction competition is now organized semi-annually, where the best structure predictors from around the world compete to determine the best structures for proteins which are currently being studied by experimental methods. This contest is known as "Critical Assessment of Structure Prediction", or CASP (Marchler-Bauer *et al*, 1997). The fourth CASP is scheduled for later this year.

There are several different methods currently being used for protein structure prediction. These include homology modeling which relies on the existence of similar, known structures, and *ab initio* methods such as protein folding. Homology modeling (Levitt, 1992) involves searching the databases of known structures and using a variety of sequence comparison techniques to find proteins which are likely to have a similar structure. The new sequence, whose structure is being predicted, is then fitted onto the structure of the known proteins, and tweaked until a reasonable new structure is obtained. This relies on the current structure databases for success, and thus is in a different category from the so called *ab initio* methods, which do not rely directly on information from currently known structures. This is an important distinction to make because even though methods which rely on database searches often perform better in real-world scenarios, the databases are not currently large enough that homologues exist for all, or even close to all proteins whose structure is unknown. This is highlighted by the fact that the CASP competition has different categories for the different types of structure prediction.

One of the *ab initio* methods is protein folding (Levitt, 1972). What this method attempts is to start with an unfolded protein chain, and then run molecular dynamics (MD) simulations on it until the protein folds into its native state. The biggest difficulty with this method is that the computer time currently available to us is not sufficient to run detailed MD simulations over a long enough time period to fold a protein of non-trivial size. Also, assuming that such computer time does eventually become available to us, it is unclear as to how the MD algorithms currently used will perform over the longer time periods. This method has gained some publicity as of late by a new initiative from IBM. Having defeated the world's greatest chess player, Gary Kasparov, with their Big Blue

machine, IBM has decided to tackle the problem of protein folding. To this end they are constructing a new computer, Blue Gene, which they estimate will have the processing power of several thousand of today's desktop computers. Even so, they predict that it will only be able to fold proteins at the rate of approximately one per year, so with the thousands upon thousands of proteins whose structures are waiting to be determined, this is clearly not a solution to the problem.

The method used here is an *ab initio* method, but differs significantly from protein folding. The method used is the Decoy Method (Samudrala *et al*, 1999a). The first step is to enumerate many possible likely conformations of a protein, given its sequence. This is what's known as generating a Decoy Set. The idea is that in this set of possible structures, only one structure can possibly have the correct conformation, and all other structures are decoys, hence the name. This Decoy Set is then evaluated with one of several scoring potentials (called potentials since many of them calculate energy or pseudo-energy values), and the best structure based on this scoring potential is chosen. There are two main problems with this method. The first is that it is difficult to generate a comprehensive Decoy Set. Levinthal's Paradox suggests that even if each residue in a protein has only 5 states, that a 100 residue protein will then have 10^{70} different states. Even if it took only 1 second to examine each state, it is clear that a random enumeration of each and every possible conformation is out of the question. Intelligent methods are thus used in order to reduce the computation necessary to generate reasonable structures. One of these methods is to use a lattice model, and place only the C-alpha carbons of the protein onto the lattice. This has the effect of both reducing the number of possible conformations due to the structure of the lattice and also restricting the protein to compact shapes typical of real protein structures due to the lattice dimensions. The other problem is with the scoring potential. Even if a very good possible structure is in a Decoy Set, there is no guarantee that a given scoring potential will be smart enough to choose it from amongst all the decoys.

Here we present a Genetic Algorithm (GA) as a method of generating Decoy Sets, and we test this algorithm on a series of small proteins whose structures have previously been determined. The Decoy Sets are then evaluated with different scoring potentials, and the results are closely examined.

- 1. Initialize the Random Population**
- 2. Add in random Secondary Structure**
- 3. Calculate Energy for each Structure**
- 4. If $E < E_{threshold}$ Structure becomes a Decoy**
- 5. Convert Energy to Fitness**
- 6. Create a Fitness Based Mating Pool**
- 7. Perform Reproduction/Crossover**
- 8. Invoke Random, Helix and Turn Mutations**
- 9. Repeat from Step#3 Until 10000 Decoys**

Figure 1: Flowchart Showing an Overview of the Genetic Algorithm

Method

The Decoy Sets used in this study are generated using a Genetic Algorithm. The GA is coded as a subroutine of the MD package Encad (Levitt *et al*, 1995). The GA subroutine interacts with Encad as follows:

1. Encad is started
2. The torsion angle minimization subroutine within Encad is called (Tormin)
3. Tormin calls the GA
4. The GA sends Tormin torsion angles one population member at a time
5. Tormin returns the calculated Energy for each population member.

The method is tested on four different proteins, identified by their Protein Data Bank (PDB) ID Number. They are:

a). 1bba	36 residues	129 torsion angles	-bovine pancreatic polypeptide
b). 1bgk	37 residues	143 torsion angles	-potassium channel inhibitor
c). 1cq0	28 residues	93 torsion angles	-hypothalamic neuropeptide
d). 1zdd	34 residues	143 torsion angles	-protein A domain

The actual algorithm used is one coded specifically for this problem, and is very similar to the GA. For a summary of some of the parameters used in this experiment please see Table 1, and for an overview of the operation of the GA please see Figure 1. A more detailed description is given here.

The Genetic Algorithm can feasibly operate on one of two features of the proteins's structure; it can operate directly on the three dimensional coordinates of the atoms in the protein, or it can operate on the torsion angles of the protein. A torsion angle is defined by any series of 4 atoms connected by three contiguous bonds (see Figure 2) and the complete three-dimensional structure of the protein can be reconstructed from the set of torsion angles for a given protein. We have chosen the torsion angle representation for this GA. The reconstruction of the structure is in this case performed by the MD package, and not by the Genetic Algorithm. We believe that there are several advantages to working within torsion angle space, as opposed to cartesian coordinate space. The first is that the different types of secondary structure of a protein are readily defined in terms of the Phi and Psi angles of the protein backbone. (these are the two torsion angles which define the backbone of a protein). The secondary structure of a protein is not however readily defined in terms of cartesian coordinates. A second advantage is conciseness. The proteins used in this experiment are all defined with less than 150 torsion angles each. However some of the proteins used are composed of upwards of 500 atoms, and each to these atoms would require three coordinates, thus a total of 1500 or more coordinate figures would be required. A third advantage lies within the simplicity of constructing the protein itself. If one constructs a protein based on a set of completely random torsion angles, the protein will undoubtedly have very poor geometry, be very unstable, and bear little resemblance to it's native structure but it will still be a contiguous set of atoms with proper bonding. However, random cartesian coordinates have no restrictions and would result in a random set of atoms in space. To keep all of the atoms in their proper bonded forms, and to keep all of the amino acids intact, and to connect all of the amino acids together requires a great deal more work in cartesian space. The representation used for the GA is thus that each member of the population is a string of real number torsion angle values, ranging from -180 degrees to +180 degrees. The main chain torsion angles (Phi and Psi, see Figure 2) are listed first in the genome, followed by the side chain torsion angles. No distinction is made between the two types of angles during the crossover event, but only backbone torsion angles are considered for helix and turn generation operators.

Representation:	Real Number String of Torsion Angles
Population Size:	2000
Fitness:	Max Energy - Energy of Individual
Decoy Selection Case:	Energy < Threshold Energy
Crossover Rate:	0.75
Mutation Rates (Standard/Helix/Turn):	0.05/1.00/0.10
Termination:	10000 Decoys are Generated

The value returned by Encad to the GA subroutine is the Encad energy of the protein. This cannot be used directly as fitness since more favorable energy is lower, and can span both positive and negative values. Fitness is thus calculated as per Equation 1.

$$(1) \quad F_n = E_{mg} - E_n$$

F_n = Fitness of Individual "n"

E_{mg} = Maximum Energy of any individual in the population for that generation

E_n = Energy of Individual "n"

The result is that fitness values range from '0' (where $E_n = E_{mg}$) to ' $E_{mg} - E_n$ ', (which can be larger than E_{mg} itself since E_n can be negative). These fitness values are then used for standard, random, fitness based reproduction using a weighted roulette wheel approach in the algorithm.

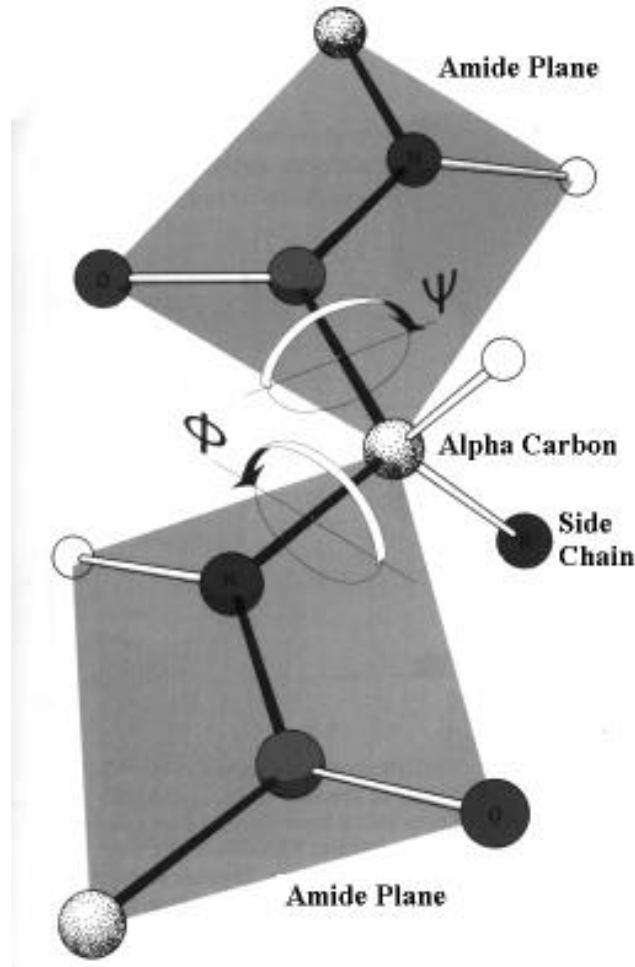


Figure 2: Here two peptide bond groups are shown, and the Phi and Psi backbone torsion angles. (Voet and Voet, 1995)

For the sake of simplicity the GA has been designed to work only with proteins that have helical or random secondary structure, but not proteins which have any significant degree of beta-stranded secondary structure. The crossover operator used is the basic crossover, but operates at two points along the chain of torsion angles instead of one. There are three mutation operators used, which operate sequentially on the population and in the order described below, after mating and crossover have occurred. The first is the standard mutational operator, which operates on each torsion angle individually, and with a small probability mutates it into another, completely random angle. The second is a helix generating operator. This operates on each individual of the population separately, and with a relatively high probability takes a series of backbone torsion angles and assigns values to them typical of alpha helical secondary structure. The helix is of random length and position in the chain, but if the length is assigned as greater than or equal to 15 residues, the operation fails and no helix is generated. The third and final mutation operator is a turn generating operator. This also operates on each individual of the population separately, but with a low level of probability. The turn operator mutates a series of four sequential backbone torsion angles, all to random values. This is designed to generate a turn or bend in the backbone of the protein, since a significant turn in a protein is defined by four torsion angles, not just one. Both the helix and turn operators affect only the backbone, not the side chain torsion angles.

The starting population is also not completely random. For each member of the population, up to five alpha helices of random length and position are generated. These are subject to the same restrictions as the helices generated with the mutational operator. One thing to note is that there is no provision for partial or complete overlapping of generated helices. What this means is that if a random helix is generated within a currently existing helix then the operation is essentially ignored, and that if two helices overlap they become one longer helix. This may seem like a relatively large number of helices given the size of the proteins, but it is representative of the large secondary structure content in most proteins.

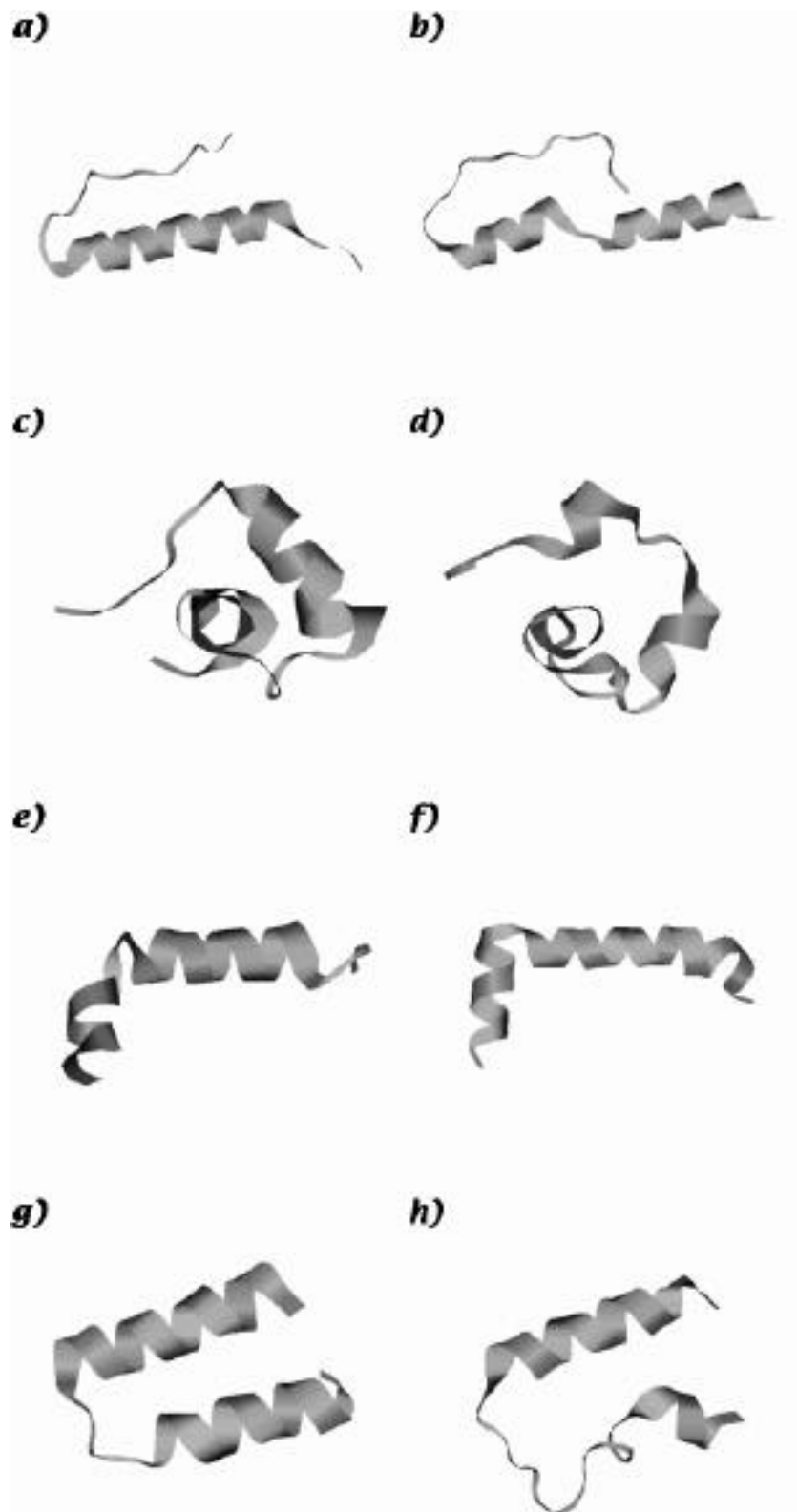
One significant difference between this GA problem and a typical GA problem is that the GA itself is not concerned with finding the best possible solution. This is the job of the scoring potential, and thus what the GA does is in fact find many possible solutions; the decoys. To determine if a given structure is accepted or rejected as a decoy, the calculated energy from Encad must be below the threshold energy. This threshold energy scales with a factor of $n^{3/2}$ with respect to the number of residues in the protein. The threshold is such that any suggested structures with an extremely bad geometry, or extremely unfavorable atomic interactions are rejected, but that structures that have any kind of reasonable conformation are accepted. The energy calculated by Encad is not however useful for discriminating between decoys to determine which is best, and thus scoring potentials are used for that purpose. The GA terminates after 10000 decoys have been generated.

Two different scoring potentials are used to evaluate the Decoy Sets. The first, Ram's Potential (Samudrala and Moulton, 1998) is an all atom-potential. That is to say it calculates the energy based on the complete atomic structure of the protein. The second, the Hmoment Potential (Samudrala *et al.*, 1999b), calculates an energy based on the clustering of the backbone carbons in the protein. One of the significant differences is that Ram's Potential will depend on the conformation and packing of the amino acid side chains, since it examines all atoms, but the Hmoment Potential will not.

Results

10000 decoys were generated for each of the four test proteins using the GA. The root mean square deviation of each of these decoys was then calculated with respect to the native structure (taken from the PDB) using the program Struatal. The RMSD was calculated using only the alpha carbons (the central carbons of each amino acid), and is reported in Angstroms. The best decoy in terms of RMSD was chosen, and is presented alongside the native structure in Figure 3. Figure 4 shows the distribution of Energy values calculated using the two potentials with respect to the RMSD for each of the 4 decoy sets. The X axis is the RMSD of a given decoy, and the Y axis is the energy as calculated by the scoring function. Both lower RMSDs and lower energies are better. The best RMSD values are given here (in Angstroms):

- a). 1bba 3.85
- b). 1bgk 4.59
- c). 1cq0 3.05
- d). 1zdd 3.27



**Figure 3: 1bba: a), b) 1bgk: c), d) 1cq0: e), f) 1zdd: g), h)
Actual Structures on the Left, Best Decoys on the Right**

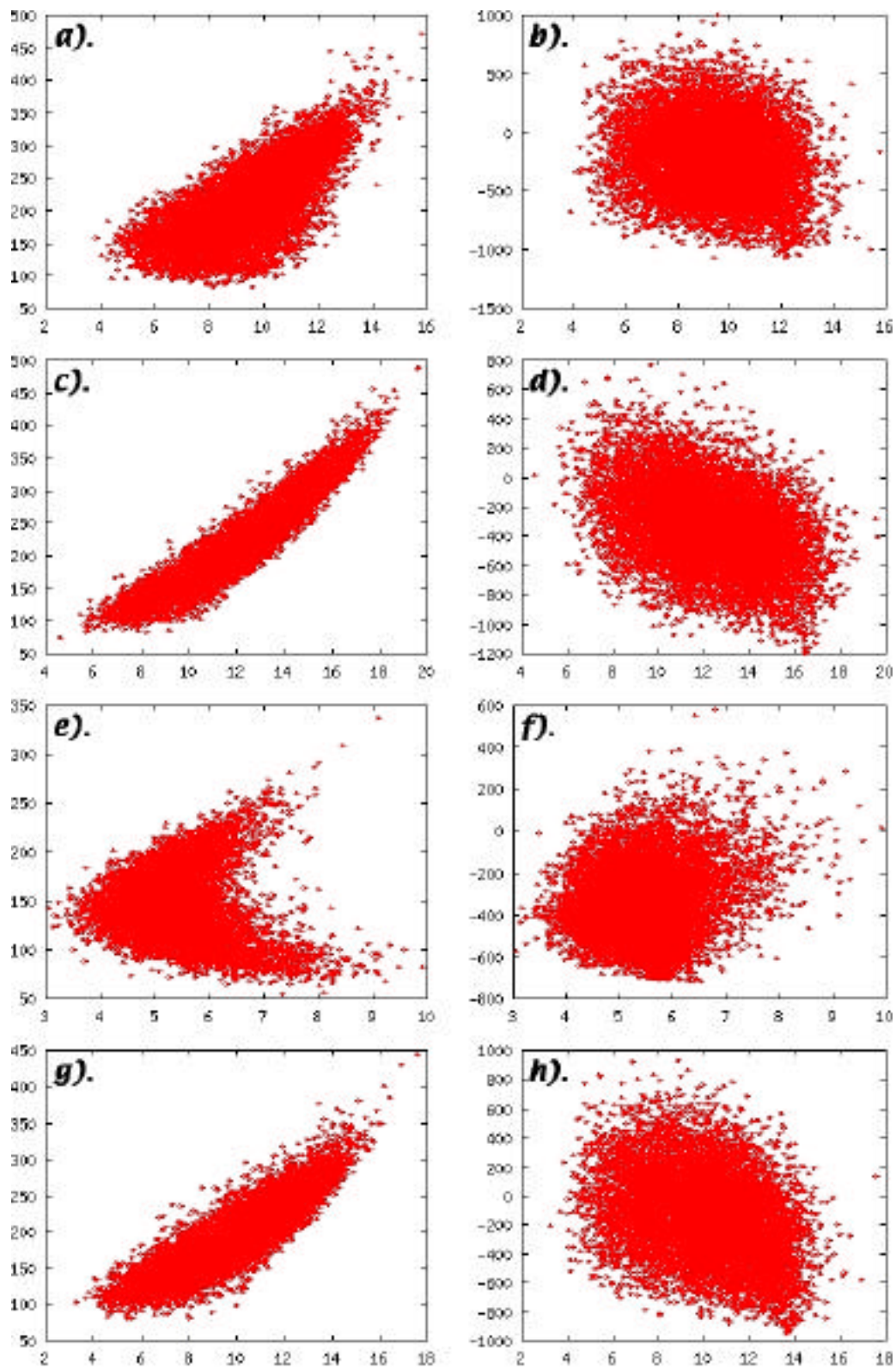


Figure 4: 1bba: a), b) 1bgk: c), d) 1cq0: e), f) 1zdd: g), h)
Hmoment Potential on the left, Ram Potential on the Right

The computer time required for the generation of the Decoy Sets is quite small. All computation was done on an Apple Powerbook G3 using a 333 MHz G3 processor running LinuxPPC 1999. Approximate times for the generation of Decoy Sets for each of the four proteins are as follows:

a). 1bba	2 hours
b). 1bgk	1 hour, 20 minutes
c). 1cq0	22 minutes
d). 1zdd	1 hour, 18 minutes

One interesting thing to note is that even though the number of generations is not used as any criterion in the experiment, the typical Decoy Set was produced after 20 -30 generations of the GA.

Discussion

The results of this method are promising, but are by no means an immediate solution to the problem of protein structure prediction. Several relatively good decoys were calculated for each of the test proteins, and the best decoy in each case does in fact capture elements of the overall topology of the protein. However, in none of the cases are the finer details of the proteins structure accurately predicted by the best decoy. There are several possible reasons for this. First of all, even though 10000 structures may seem like a rather large number, the GA used in no way accounts for there being similar structures, or even exact copies in the Decoy Set. With the high crossover rate and high combined mutation rates it is unlikely that too many exact duplicates occurred, but it is the nature of the GA that similar members show up in the population from one generation to another. What this means is that the true number of distinct decoys may in fact be significantly lower than 10000, and that a good sampling of the conformation space is not obtained. Further, the secondary structure is handled in a very primitive manner. Even though it was earlier acknowledged that the GA was purposely limited to proteins with only alpha helical secondary structure, the alpha helix is not an exact entity. The Phi and Psi values used in the generation of alpha helices by the GA are -57 and -40 degrees respectively, but in real structures these values can vary by as much as 10 degrees on either side. Also, given the extremely large search space of protein structure, a population size of 2000 is by no means large. It is possible that a larger population size would give a wider range of initial structures, and thus a much better distribution of structures within the Decoy Set.

Another issue is that of the ability of the scoring functions to distinguish between good and bad decoys with respect to their RMSD values. In no case does Ram's Function reveal a positive correlation between RMSD and Energy, in fact it tends towards a negative correlation in Figure 4d) and 4h). One reason for this is that the selection of decoys by the GA is rather lenient, requiring only that the structures fall under a rather generous energy value (as calculated by Encad). This allows for the protein side chains to be quite poorly packed, and also allows for atoms to infringe on the physical space of other atoms. Due to the generation of secondary structure elements (which have a good physical geometry pre-defined), this problem is not so apparent with the backbone atoms, but is quite apparent with the side chain atoms. Since Ram's Potential considers all atoms, it penalizes the side chain clashes heavily. This results in a negative correlation since the best structures in terms of RMSD are usually amongst the most compact, but the more compact structures also have more overlapping side chains, and thus a higher energy.

The Hmoment Potential does not however consider side chain atoms, and only considers the carbon atoms in the backbone of the protein. What this means is that the more compact, and more realistic structures are more likely to exhibit better energies with the Hmoment Potential, even though they may suffer from poor side chain structure. A solution to this problem is to build better side chains in the Decoy Set. One way to accomplish this would be to perform an energy minimization routine on the side chain atoms only for each decoy, allowing them to pack more favorably with each other, and avoid the large energetically unfavorable interactions that are currently occurring. This could be done during decoy generation, or afterwards with the Encad program. This would undoubtedly provide better structures, but would also drastically increase the computation time. Given that the computation time is currently quite small it is a feasible approach. Another approach would be to only build the main chain of the protein with the GA. The Decoy Set would initially consist of only backbone atoms, and then the side chains would be built in separately.

An interesting result is Figure 4d), the RMSD vs Energy for 1cq0 calculated using the Hmoment Potential. The graph here is bimodal, indicating two separate relationships. Looking at the structure of 1cq0 we can provide an explanation for this behavior. The structure of 1cq0 can be viewed in one of two ways; it is either two alpha helices,

or one alpha helix bent at about 90 degrees. If one supposes that it is one bent helix, the structure does not need to be altered very much to straighten out the helix. Now consider the Hmoment Potential; it ranks structures based on the compactness and clustering of main chain carbon atoms. Both the bent and straight helix structures are similarly compact, and have similar clustering of atoms. Thus they would both have low energy when ranked with Hmoment. However, only the bent helix has a good RMSD. Thus the second graph mode, which is horizontal and spans from low to high RMSD is due to the decoys which are straight helices, and which therefore have low energies.

The case of 1cq0 aside however, the Hmoment Potential does in fact demonstrate a strong correlation between the Energy of the decoy and the RMSD value, and as such proves useful in choosing the lowest RMSD structure from the Decoy Set. A rather exceptional case is that of 1bgk (Figure 4c), where the best decoy (4.59 Angstrom RMSD) also exhibits the lowest energy of any of the decoys. This is the ideal scenario for this method of protein structure prediction and reveals great promise for the method. Figure 4g also shows a good correlation, and the Hmoment Potential would select a 4.5 Angstrom structure, though the best structure was just over a 3 Angstrom RMSD.

Conclusions

The GA used here is rather simple, and can undoubtedly benefit from many improvements. We have already touched upon its inability to handle beta-strand secondary structure and the poor construction of side chains. There are also the issues of population size and mutation rates, which need to be explored more thoroughly, and the performance of the GA is unknown for proteins of larger size. These caveats aside, the GA does remarkably well in predicting decoys with a reasonable RMSD, especially considering the small computational times required. Refinement of the algorithm, the resolution of the aforementioned issues, and the addition of new features such as amino acid specific information could greatly increase the GA's performance in this field.

Acknowledgements

I would like to thank Michael Levitt and Jody Puglisi for the computer used to run the experiments, and Michael Levitt for providing the source code for both the Encad and Structural programs. Ram Samudrala is responsible for both the Ram and Hmoment Potentials. I would also like to thank Yu Xia for help in running the scoring potentials, and Golan Yona for code debugging.

References

- Levitt, M. 1972. Folding of Nucleic Acids. In Polymerization in Biological Systems, *Ciba Foundation Symposium* 7, Elsevier, Amsterdam, 146-171.
- Levitt, M. 1992. Accurate Modelling of Protein Conformation by Automatic Segment Matching. *J. Mol. Biol.* **226**: 507-533.
- Levitt, M., M. Hirshberg, R. Sharon and V. Daggett. 1995. Potential Energy Function and Parameters for Simulations of the Molecular Dynamics of Proteins and Nucleic Acids in Solution. *Computer Physics Communications.* **91**: 215-231.
- Marchler-Bauer, A., Levitt, M and S. Bryant. 1997. A Retrospective Analysis of the CASP2 Threading Predictions. *Proteins, Struct., Funct. and Gen. Suppl.* **1**: 83-91.
- Samudrala R, Moulton J. 1998. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *Journal of Molecular Biology* **275**: 893-914.
- Samudrala, R., Y. Xia, E.S. Huang, and M. Levitt. 1999a. Bona Fide Ab Initio Protein Structure Prediction Using a Combined Hierarchical Approach. *Proteins, Struct., Funct. and Gen. Suppl.* **3S**: 194-198.
- Samudrala R, Xia Y, Levitt M, Huang ES. 1999b. A combined approach for ab initio construction of low resolution protein tertiary structures from sequence. *Proc. of the Pacific Symposium on Biocomputing*, 505-516.
- Voet, D. and Voet, J. 1995. *Biochemistry 2nd Edition*, John Wiley & Sons, New York.