Evolving an Improved Axial Structure for Fibrillar Collagen

D.E. Cairns
Dept of Computing Science and Mathematics
University of Stirling
Stirling, U.K.
FK9 0QL
+44 (0) 1786 467445
dec@cs.stir.ac.uk

G.J. Cameron
Dept of Computing Science and Mathematics
University of Stirling
Stirling, U.K.
FK9 0QL
+44 (0) 1786 467447
gjc@cs.stir.ac.uk

T.J. Wess
School of Optometry and Vision Science
Redwood Building
Cardiff University
Cardiff, CF10 3NB
+44 (0) 29 2087 0117
wesstj@cardiff.ac.uk

Categories & Subject Descriptors: I.2.8 Problem Solving, Control Methods, and Search - Heuristic methods; J.3 Life and Medical Sciences - Biology and Genetics

General Terms: Algorithms

Keywords: Bioinformatics; Genetic Algorithm; Collagen; X-ray Diffraction

1. INTRODUCTION

In this study, we investigate the modeling of non-uniform axial translation of amino-acids in the collagen helix through the use of a Genetic Algorithm [1][2][3]. The aim of the study was to evolve a molecular structure that produced a simulated diffraction pattern with best fit to observed X-ray diffraction data.

Determination of fibrillar collagen structure has to date relied mostly on X-ray diffraction and electron microscopic evidence. In both cases the broad structural features can be accounted for using models where the amino acid sequence is translated into electron density by placing amino acids at regular co-ordinate locations as dictated by the collagen helix. However studies that seek to elucidate structures at higher resolution fail to maintain the correlation between the predicted position of amino acids and the electron density profile, this indicates regions of possible rarefaction and compaction of electron density in the fibrillar structure.

Collagen is a complex structure. In terms of representing this structure for a GA, there are a number of key concepts that need to be considered such that the simulated output from a Fourier inversion process applied to the model produces factors that relate to the observed data. For the purposes of this study, the key factors we have identified are the core triple helix, the telopeptides at either end of the helix and fold points in the telopeptides.

2. METHOD

In this study, the parameter set for the genetic algorithm comprised of individual amino acid spacing factors that were assigned to each amino acid type and used to generate spacing in the molecular structure of the triple-helix and telopeptide regions, together with fold points in the telopeptide region.

The main parameter affected using this approach is the inter-amino acid spacing and therefore the most logical approach would be to assign individual values for each amino acid pair in the three chains. However, this would produce a model with a high degree of freedom and possible inconsistencies. Using this particular approach, it would be possible for two pairs of identical longitudinally adjacent amino acids to have a different inter-amino acid spacing at different places on the same chain.

The approach chosen in this study was to assign a factor to each of the 20 amino acid types present in the collagen sequence and then use this factor to determine the relative size of the distance for a given amino acid pairing. The task of the GA was to modify these factors in such a way that a close fit to the observed intensity data was achieved. These factors provided significant constraints upon the degrees of freedom in the model, since any adaptation to the factor of one amino acid would result in multiple changes throughout the model structure.

In order to constrain the model and ensure that the amino-acids in each of the chains remained in step as one progressed down the collagen molecule, it was decided that the amino acids in the three chains should be grouped into 'lateral' triplets with each of the three chains contributing one amino acid. The GA calculated the distance between these triplets as follows.

$$S_{An,An+1} = \left(\frac{A_{n,1}.A_{n+1,1}}{0.286} + \frac{A_{n,2}.A_{n+1,2}}{0.286} + \frac{A_{n,3}.A_{n+1,3}}{0.286} \right) / 3$$

In this approach the multiplied factors are scaled using 0.286 in order to enforce normalization upon them such that they remained within specific constraints.

It should be noted that the factors calculated in the methods shown above are not directly related to the actual distances between amino acids within the collagen molecule. Each assigned factor should be examined relative to the factors assigned to the other amino acid types in order to provide information on the nature of the relationship between the different amino acid pairings.

2.1 Fitness Function

Given the biological constraints on the model, it was important that the fitness function properly accounted for relevant factors such that a model could be obtained that gave both a positive match to the observed X-ray diffraction data and was also biologically feasible. The fitness function should however allow the GA to produce solutions with a strong match to the observed

GECCO'05, June 25-29, 2005, Washington, DC, USA.
ACM 1-59593-010-8/05/0006.

data but relatively weak biological feasibility if such structures could then be improved upon from a biological viewpoint.

Rather than discarding a model that had fallen outside biological constraints, it was only penalized but left in the population pool. This allowed a traversal of weak solution space in order to enable the GA to locate solutions with both a high fit to the observed data and that were also biologically correct.

The authors had to spend some time fine tuning this balance since hard limiting the biological feasibility diminished the ability of the GA to find any solutions at all. Reducing the relevant weighting factors eventually led to the evolution of solutions that matched both sets of criteria.

In this study, a number of measured objectives were considered when calculating the fitness function. These included:

1. The level of fit between the observed intensity and the equivalent data from the model.
2. The total length of the simulated molecule.
3. The maximum and minimum values for the inter-amino acid distance and the average inter-amino acid distance present within the helical part of the simulated molecule.

A score, *S*, is derived by taking the sum of the squares of the weighted differences between the model and the observed data for each of the amino acids and biological factors.

$$S = \frac{\sqrt{\sum_{n=1}^{N} (O_n - M_n)^2 \, W_n}}{C}$$

This is the standard weighted *N* dimensional Euclidean distance between the intensities in the solution set where *N* is the number of factors in the model, W_n is the weight assigned to each of the factors under study, O_n is the required value of factor *n* and M_n is the value of factor *n* produced by the model. C is a normalisation component defined as the maximum value that the upper term can take. A further validation of the model under scrutiny is to calculate the R-factor.

$$Rfactor = \sqrt{\frac{\sum_n \left(O_n - \left(\left(\frac{\sum_n O_n . \sum_n M_n}{\sum_n (M_n^2)} \right) . M_n \right) \right)^2}{\sum_n (O_n^2)}}$$

In the case of each proposed structure, the resultant molecular model was tested for structural feasibility against known criteria. In order to generate a constrained set of models, the GA penalized model structures that did not satisfy the relevant criteria. The criteria used were features that are readily recognized as being inherent to collagen molecules, with the most heavily penalized aspect being the length of the molecule. This is 300 nm ± 3[5].

Multiple runs of the GA were performed with a population size of 50 candidate solutions. Each run was allowed to evolve until the change in fitness score had reached an asymptote that was within 0.001% of the fitness scale. This typically was on the order of 400,000 epochs per run. Based on the results of each run, the weightings for relevant biological factors such as molecule length were adjusted until the GA delivered more feasible results.

3. RESULTS

The best fit to the observed data when scored over the first 30 orders of intensity taken from observed X-ray diffraction data is shown below. This model achieved an R-factor value of 0.049 which is a significant improvement upon models that have been generated by using alternate techniques such as simulated annealing [4].

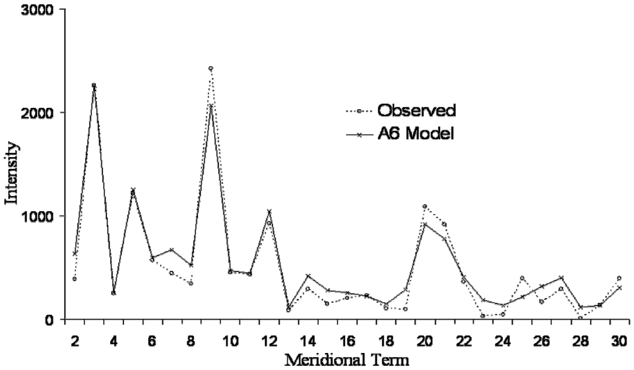

Figure 1 : Model scored over 30 orders of intensity. The R-factor for this model was recorded as 0.049.

4. CONCLUSION

This study has demonstrated how GAs can be adapted to tackle the problem of amino acid positioning with the triple helical region of a collagen molecule. Although we have demonstrated the principle of fitting a GA evolved molecular structural model to X-ray diffraction data for collagen, this approach should be transferable in order to determine the structure of a broader set of protein molecules.

5. REFERENCES

[1] Fogel G B, Corne D W. (eds), Evolutionary computation in bioinformatics. (2003), Elsevier Science, San Francisco.

[2] Goldberg D E. Genetic Algorithms in Search, Optimization & Machine Learning. (1989) Addison Wesley.

[3] Holland J H. Adaptation in Natural and Artificial Systems. (1975) University of Michigan Press. (2nd Ed: MIT Press 1992).

[4] Orgel J P R, Wess T J, Miller A. The in situ conformation and axial location of the intermolecular cross-linking non-helical telopeptides of type I collagen. (2000) Structure 8 137-142.

[5] Schmitt F O, Gross J, Highberger J H. Tropocollagen and the properties of fibrous collagen. (1955) Experimental Cell Research, Supplement 3 326-334