NK Landscape Instances Mimicking the Protein Inverse Folding Problem Towards Future Benchmarks

Sune S. Nielsen
CSC Research Unit,
University of Luxembourg
6, rue Coudenhove-Kalergi
L-1359 Luxembourg
sune.nielsen@uni.lu

Grégoire Danoy
CSC Research Unit,
University of Luxembourg
6, rue Coudenhove-Kalergi
L-1359 Luxembourg
gregoire.danoy@uni.lu

Pascal Bouvry
CSC Research Unit,
University of Luxembourg
6, rue Coudenhove-Kalergi
L-1359 Luxembourg
pascal.bouvry@uni.lu

El-Ghazali Talbi
INRIA Lille Nord Europe,
Villeneuve d'Ascq, France
el-ghazali.talbi@inria.fr

ABSTRACT

This paper introduces two new *nominal NK Landscape* model instances designed to mimic the properties of one challenging optimisation problem from biology: the Inverse Folding Problem (IFP), here focusing on a simpler secondary structure version. Through landscape analysis tests, numerous problem properties are identified and used to parameterise and validate model instances in terms of epistatic links, adaptive- and random walk characteristics. Then the performance of different Genetic Algorithms (GAs) is compared on both the new *NK Models* and the original IFP, in terms of population diversity, solution quality and convergence characteristics. It is demonstrated that very similar properties are captured in all presented tests with a significantly faster evaluation time compared to the real IFP. The future purpose of such a model is to provide a generic benchmark for algorithms targeting protein sequence optimisation, specifically in protein design. It may also provide the foundation for more in-depth studies of the size, shape and characteristics of the solution space of good solutions to the IFP.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

Keywords

NK Landscape; Landscape analysis; Genetic Algorithm; Benchmark function

GECCO '15, July 11 - 15, 2015, Madrid, Spain

DOI: http://dx.doi.org/10.1145/2739482.2768438

1. INTRODUCTION

Protein structure prediction is an essential step in understanding the molecular mechanisms of living cells with widespread application in biotechnology and health. Conventional protein folding prediction research is concerned with finding or predicting the folded structure of a given amino acid sequence. To the present day the problem is not solved but scientists have early on sought to simplify it by solving the inverse problem, referred to as the Inverse Folding Problem (IFP). The latter consists in finding sequences that fold into a defined structure. The IFP is an important research problem that is at the heart of most rational protein design approaches.

Due to its quickly exploding complexity and highly multimodal nature, it is a challenging task to determine all or a fraction of its local optima. In addition, tackling real biological instances is computationally very expensive which therefore limits the number of possible experiments.

In this work some of the problem characteristics are sought identified to design a model that captures the most prominent of these. With a simple definition based on the well-known *NK Model*, the motivation is to make the IFP problem more accessible to algorithm specialists and model experts contrary to being a problem solved mostly by bioinformaticians with main expertise in other fields.

The remainder of this article is organized as follows. The next section introduces the related work on the NK Model and on the inverse folding problem. Then the proposed NK model is presented in detail in section 3, followed by the landscape analysis of the original problem and its comparison to the proposed NK model in section 4. Finally conclusion and perspectives are provided in section 5.

2. RELATED WORK

This section presents a few relevant works related to the two main areas covered in this work: The NK Model and the Inverse Folding Problem model based on secondary structure prediction.

2.1 The NK Model

The *NK Model* introduced by Kaufmann [2] is a tunable rugged fitness function designed to model complex epistatic links among variables, to study topics such as gene-interaction. A central feature of the model is its stochastic design which opens up possibilities for statistical analysis of its properties without exact knowledge of all underlying epistatic interactions. While the original model works on a bit-string encoding, Li *et al.* extended the model to continuous and mixed integer solution spaces [3]. Specifically the *nominal discrete NKL model* is of interest, where L denotes the possible values at each allele location with $L = 2$ defining the binary case corresponding to the original *NK Model*. The original *NKL Model* is described in Equation 1 which implies that any allele x_i and its K neighbors x_{i1}, x_{ik} contribute to the function value.

$$F_{NKL}(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} F_i(x_i; x_{i1}, ..., x_{ik}), \mathbf{x} \in \{0, L\}^N \qquad (1)$$

Most common neighborhoods are defined by the K adjacent positions left and right from the position i or K random positions in addition to i, making $K = N - 1$ the maximum possible value for K. Typically the model is made circular to avoid boundary effects.

2.2 Protein Inverse folding Problem

The structure of a protein can be divided into different levels (see Fig. 1). The primary structure is the protein sequence of N amino acids $\{aa_i\}$ where $1 \leq i \leq N$ is the residue position. The secondary structure defines or annotates the organisation of *helices*, *sheets* and *loops* of the tertiary structure and can be expressed by a type $\{T_i\} \in \{H, E, L\}$ for each position i in the protein. The tertiary structure completely describes the arrangement of all atoms of a single sequence in three-dimensional space.

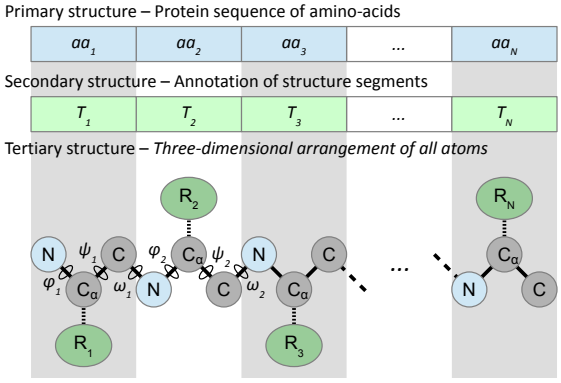


Figure 1: Three levels of protein structure

A protein sequence determines the structure of a protein, however a given structure can be obtained by *more* than one protein sequence. The solution to an instance of the protein Inverse Folding Problem (IFP) is ideally the complete set of all sequences that fold into the given reference protein structure. In practice researchers focus on finding a limited number of new matching sequences which are as different as possible from the original reference sequence. The inverse folding problem has been tackled in [4] where a simplified model was developed to matching solely the reference secondary structure - a requirement for the tertiary structure to match. This is motivated by the fact that computing the tertiary, i.e. 3D, structure of a given input sequence is computationally very expensive, which prevents the usage of a metaheuristic on the entire sequence. Using the PROFphd tool, updated to ReProf [5], the likely secondary structure type $T_{pred}(i)$ can be predicted per amino acid aa_i in A with a reliability, $R_{pred}(i) \in \{0...9\}$ by means of posterior neural network training. With $T_{ref}(i)$ the actual type found at position i of the reference secondary structure, the estimated similarity score $F_{sec}(A)$ is calculated as a sum of reliability weighted (mis)matches:

$$F_{sec}(A) = -\frac{\sum_{i=1}^{N} s_i \cdot (C^R_{pred} + R_{pred}(i))}{\Sigma_{max}} . \qquad (2)$$

where

$$s_i = \begin{cases} 1 & \text{if } T_{pred}(i) = T_{ref}(i) \\ -1 & \text{if } T_{pred}(i) \neq T_{ref}(i) \end{cases}$$

and

$$\Sigma_{max} = (C^R_{pred} + \max R_{pred}) \cdot N$$

C^R_{pred} is a constant which purpose is to increase the contribution to the score of a matching type prediction that has a low reliability R_{pred}. In the current work it was chosen such that $C^R_{pred} + \max R_{pred} = 20$. By using the objective function $F_{sec}(A)$ in Equation 2 as target for optimisation algorithms, likely solutions to the IFP have been found [4]. The next section essentially presents an alternative lightweight function designed to have very similar properties.

3. PROPOSED NKL MODEL

The proposed *NK model* is presented in Equation 3, which is a variation of Equation 1. It omits the contribution of the i_{th} position in $\mathbf{x}$, hence K for an identical neighborhood will be one larger than in the original model and the maximum K becomes $K = N$. This is a minor change that allows to re-create epistatic link effects of the target IFP problem. In addition, the model uses a single function F_0 instead of N different F_i functions. This is for simplicity reasons as N may exceed values of 100 and in theory two random functions based on the same distribution are equivalent. $N = 67$ is chosen because the actual sequence of the target IFP protein *1b3a* has length 67. Then by fixing the number of nominally discrete values possible at each allele position to $L = 20$, a solution vector $\mathbf{x} = \{x_i\},\ x_i \in \{1...20\}$ for the model can be translated 1:1 from an RNA sequence $A = \{aa_i\},\ aa_i \in \{1...20\}$ of the 20 possible amino-acids. This effectively makes the solution encoding of the model and the IFP identical seen from the point of view of an algorithm or solver. Hence, an algorithm designed to work with amino-acid sequences can easily be adapted to solve the proposed model and vice-versa.

$$F(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} F_0(x_{i1}, ..., x_{ik}), \mathbf{x} \in \{0, L\}^N \qquad (3)$$

The novelty in this work is the combination of two *NK Models*, $F^A(\mathbf{x})$ and $F^B(\mathbf{x})$, with different K and different neighborhood definitions by a simple multiplication:

$$F(\mathbf{x}) = F^A(\mathbf{x}) \cdot F^B(\mathbf{x})$$

With this setup, the combined model $F(\mathbf{x})$ can accumulate the characteristics of both its underlying models. Say, strong epistatic interactions are observed between alleles i and j in F^A as well as between k and l in F^B. The combined model will then show interactions for *both* pairs i and j as well as k and l. The objective of this setup is ultimately to come as close as possible to the original IFP that features both strong epistatic interactions between close alleles, and a constant interaction between alleles farther apart.

Two novel *NK Model* instances have been created with the following settings:

- Model 1
  - F^A: a $K = 4$ semi-adjacent circular neighborhood is designed as follows: $\{x_{i-2}, x_{i-1}, x_{i+1}, x_{i+2}\}$, omitting the central position x_i.
  - F^B: a $K = 3$ neighborhood of uniform random distribution.
- Model 2
  - F^A: a $K = 4$ neighborhood as Model 1.
  - F^B: a $K = 5$ neighborhood of uniform random + 20 positions wide triangular distribution.

The purpose of using a triangle distribution in Model 2 is to induce a higher linkage between alleles closer to each other. Essentially the chance of linking two alleles drops off linearly to ± 10 alleles apart and is then constant. The effects of the presented neighborhoods used in F^A and F^B on epistatic linkage is seen in Figure 5 and discussed further in the following section.

4. LANDSCAPE ANALYSIS

With the introduction of the *NK Model* [2], a number of model features were analysed, mainly by characterising adaptive and random walks in the landscape. Analysis of epistatic links among model variables is another important way of characterising a problem instance, which will be described in the following. As protein sample for the IFP, only *1b3a* is considered as previous work [4] has suggested that different protein samples show very similar characteristics.

4.1 Adaptive Walks

An adaptive walk starts at a random position in the objective space and progresses by choosing random *1-mutant* fitter neighbors until no fitter neighbors can be found, and a *local optimum* has been reached. This provides several indicators on the landscape, including the length of such walks and how the number of fitter neighbors decreases with each step. From literature it is known that the length of an adaptive walk on a *NK Landscape* will decrease for larger K values regardless of the choice of neighborhood. This is due to the induced ruggedness when using larger K. The effect can be seen in Table 1 for the standard *NKL Models* with $K \in \{3, 4, 5\}$. Other models in the table include the actual *IFP* objective function and the two combined *NK Model* variants proposed in this paper, averaged over 100 individual tests for each. It can be seen that the effect of combining two *NK Models* increases the length of a walk approaching that of the IFP for *Model 1*. All models show almost the same average number of fitter neighbors at the first step, ± 636.5, which is exactly half of the neighborhood size of $N \cdot (L-1) = 67 \cdot 19 = 1273$. This number shows higher variation in the *IFP* problem, indicating more location-dependent characteristics than those expressed in the *NKL* models. The number of evaluations required on average to reach a local optimum is a bit higher for the *NK Models*, and the deviation of fitness values at such optima is slightly higher for the *IFP*, though *Model 1* comes close with 0.019 vs 0.023. All in all the *NKL Model* statistics can roughly be fitted within a maximum factor of two of the *IFP* problem, and in most cases a far better match is achieved.

4.2 Random Walks

To compute the auto-correlation function of the problem and models, a number of random walks have been performed starting from a local optimum. The reason for choosing a local optimum as a starting point is motivated by the fact that the main dynamics of the estimated secondary similarity score $F_{sec}()$ are present only when the predicted structure matches the reference structure well. Evolutionary algorithms will mostly be evolving around such good solutions, and small perturbations in key positions here have larger impact on the overall match score than in a random poor matching solution. The auto-correlation function used in this work is equivalent of the one in [1] and can be written as co-variance of function values at t and $t+s$ over the product of their deviations.

$$R(t,s) = \frac{\sigma(F(\mathbf{x}_t, F(\mathbf{x}_{t+s}))}{\sigma(F(\mathbf{x}_t)) \cdot \sigma(F(\mathbf{x}_{t+s}))} \tag{4}$$

As the walks all start from local optima, the analysis will use $t = 1$, and analyse decay in correlation as the hamming-

Model	Walk length	Fitter, first step	Average fitter	Final evaluations	Final fitness
IFP	111.070 (15.811)	633.117 (160.977)	234.069 (189.727)	4896.210 (2070.319)	-0.899 (0.023)
$Model1$	95.750 (10.629)	633.330 (79.704)	161.774 (180.554)	7649.454 (2440.165)	-0.659 (0.019)
$Model2$	83.397 (9.520)	641.974 (93.780)	154.469 (181.191)	7734.680 (2797.414)	-0.633 (0.017)
$K = 3$	93.290 (9.659)	620.210 (75.056)	167.943 (178.849)	6568.108 (1561.485)	-0.896 (0.011)
$K = 4$	75.030 (9.598)	618.480 (88.863)	164.381 (180.239)	5915.798 (2011.225)	-0.869 (0.010)
$K = 5$	66.000 (8.464)	645.500 (90.636)	161.252 (184.934)	5684.346 (1521.157)	-0.850 (0.012)

Table 1: Adaptive walk statistics

distance s increases. The random walks were repeated 100 times from different local optima for the IFP and the two proposed models with the average auto-correlation shown in Figure 2.

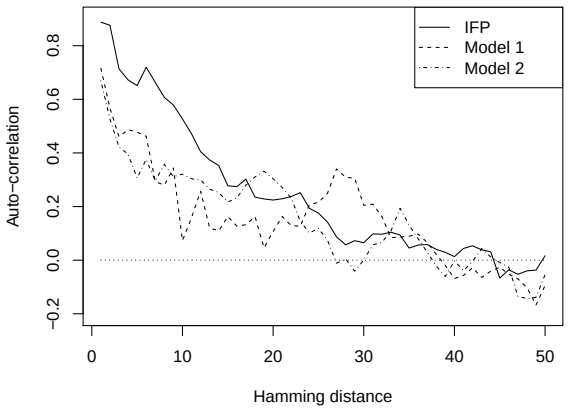


Figure 2: Auto-correlation of random walks starting from local optima.

Though the actual correlation lengths are 45, 39 and 27, for IFP, Model 1 and 2 respectively, they all seem to reach zero correlation at the same distance of about 40 in general. The shorter model correlation lengths can be explained by the higher variation in correlation, and the slightly faster decay than the IFP at shorter hammin-distances. Overall, the decay in correlation as the distance increases follows a quite similar pattern.

4.3 Epistatic link analysis

Epistatic interaction is a concept borrowed from genetics where two genes can be defined as being epistatically linked if the effect of one depends on the state of the other. To fully discover such links would require to observe the effect of all possible combinations of two genes in all possible states of all other genes. In this analysis of epistatic links, alleles of a solution are examined pairwise in a systematical manner, keeping all other genes constant. Again a local optimum is chosen as the starting point. For two selected alleles i and j, $i \neq j$, three additional function evaluations are done evaluating first a mutation at i, then a mutation at j computing the error $\varepsilon(\mathbf{x}, i, j)$ by comparing to the same mutations at both i and j at the same time:

$$\varepsilon(\mathbf{x}, i, j) = |\Delta F(\mathbf{x}_{(i,j)}) - (\Delta F(\mathbf{x}_{(i)}) + \Delta F(\mathbf{x}_{(j)}))|$$

Where $\Delta F(\mathbf{x}_{(\mathbf{y})})$ denotes the function value difference in F when the solution $\mathbf{x}$ has values substituted at allele locations $\mathbf{y}$. If there is no linkage between alleles i and j at location $\mathbf{x}$, $\varepsilon(\mathbf{x}, i, j)$ will be *zero* for all possible substitution pairs. This information is typically expressed on matrix form, but reduced here to a single vector, averaging the linkage in terms of allele distance $d = |i - j|, i \neq j$. Figure 4(a) and (b) show this epistatic linkage at two different random local optima of the *IFP* problem. Figure 5(a) and (b) shows

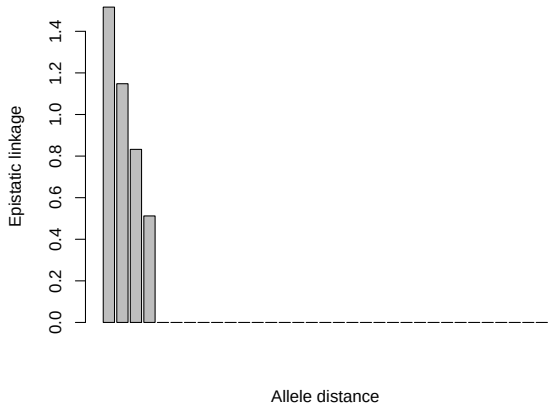


Figure 3: Epistatic linkage in local optima of NKL Model.

linkage of the proposed models at a local optimum and Figure 3 the standard *NKL Model* with $K = 5$ for comparison.

Clearly the standard *NKL Model* has absolutely no linkage beyond 5 loci apart, which is achieved in the combined models proposed here with the second function F^B having almost uniformly distributed neighborhood. To achieve the ramp down which can be observed in the real *IFP* problem, the neighborhood of function F^B of *Model 2* is generated from a partially triangular distribution, which effect is quite noticeable in Figure 5(b). The epistatic links are slightly stronger between close alleles in the models than in the IFP but long range interactions look very similar in both models. The other main feature of the real *IFP* problem is the characteristic dip and then rise in locations 2 and 3 apart which is captured by the neighborhood function of F^A and observed in both combined models in Figure 5(a) and (b).

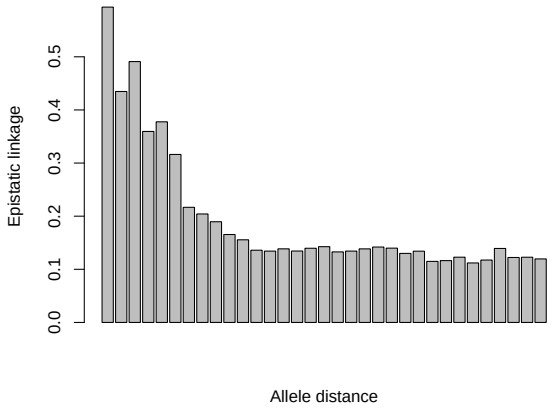
4.4 Evolutionary algorithm analysis

As a final comparison, the two *NKL Models* and the *IFP* are solved with a standard Genetic Algorithm (GA) and the NSGAII with Diversity-as-Objective and Quantile Constraint
(NSGAII-DAO-QC) algorithm proposed in [4]. The main feature of the latter algorithm is to maintain a high and controllable degree of diversity which allows studying the exploration-exploitation trade-off.

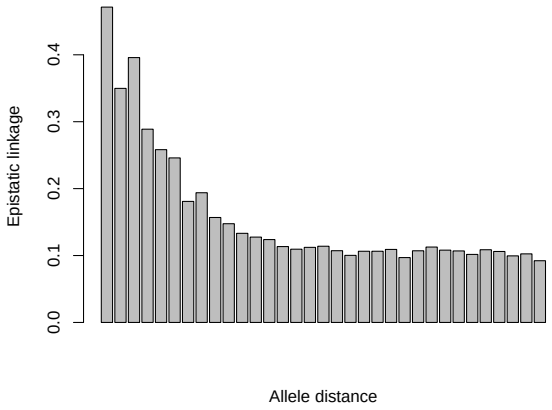
Table 2: Algorithm settings

Setting	*Value*
Population size	100
Algorithm	NSGA-II and std GA
Termination condition	30000 function evaluations
Selection	Binary tournament (BT)
Crossover operator	1-point, p_c=1.0
Mutation operator	Uniform, $p_m = \frac{1}{N}$
Quantile constraint	$C_q \in \{0\%, 5\%, 10\%, 25\%\}$

Table 2 summarises the settings: Both algorithms use a population of 100 individuals, a binary tournament selec-

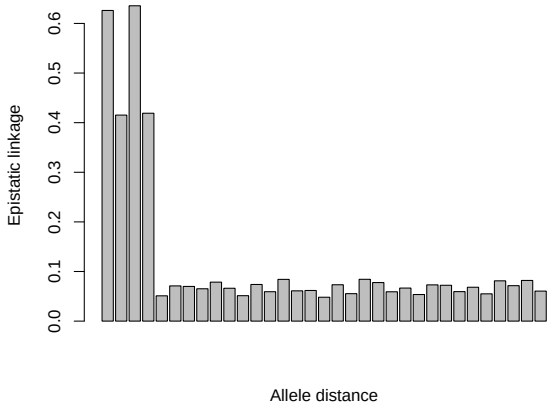


(a) Local optimum 1

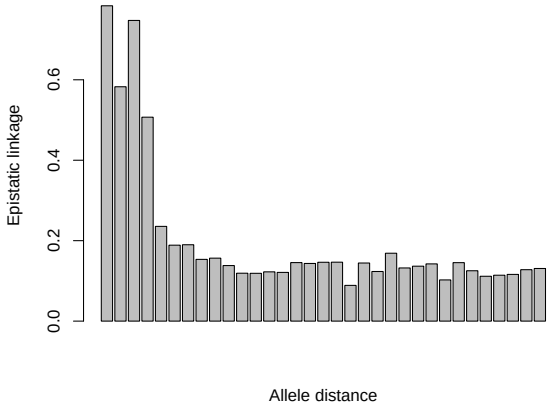


(b) Local optimum 2

Figure 4: Epistatic linkage in two local optima of protein *1b3a*.



(a) Model 1



(b) Model 2

Figure 5: Epistatic linkage in local optima of proposed models.

tion, 1-point crossover with probability p_c=1.0 and uniform mutation with probability $p_m = \frac{1}{N}$. The termination condition was set to 30000 fitness function evaluations and each experiment was repeated 30 times. Four different values of the quantile constraint C_q are considered: 0%, 5%, 10% and 25% of the population resulting in less-to-more exploitation and more-to-less population diversity, hence more-to-less exploration.

Figures 6 and 7 show convergence of fitness and diversity for the protein *1b3a*, Figures 8 and 9 the same for the *NK Model 2*. Model 1 has been omitted as differences between the models are minimal and not relevant here. Overall the ordering of the series for different algorithms and settings is strikingly similar, especially in the diversity plots. The impact of diversity on the fitness function is more significant in the models than in the IFP which probably explains why the GA has better performance in the IFP than in the models.

The information in the convergence plots is supported by the pairwise comparisons of the algorithms mean values difference in Tables 3, 4, 5 and 6. The Wilcoxon test indicator [7] with a 5% significance level provides statistical confidence in comparing the sets with symbols '▲', '▽' and '-' indicating superior, inferior and no difference. In terms of fitness, the algorithms are ordered in the following way: $QC25 = QC10 > QC5 > GA > QC0$ and $QC25 > QC10 > QC5 > GA > QC0$ with statistical confidence for protein *1b3a* and *Model 2* respectively. In terms of diversity, the order becomes $QC0 > QC5 > QC10 > GA > QC25$ and $QC0 > QC5 > QC10 > QC25 = GA$. These minor differences can be explained by the difference in sensitivity of fitness to diversity mentioned earlier.

	GA	DAO-QC0	DAO-QC5	DAO-QC10	DAO-QC25
GA	⁄	−0.272▽	0.0126▲	0.0432▲	0.0541▲
DAO-QC0		⁄	0.285▲	0.316▲	0.327▲
DAO-QC5			⁄	0.0306▲	0.0415▲
DAO-QC10				⁄	0.0109 -
DAO-QC25					⁄

Table 3: Protein *1b3a* average fitness delta

	GA	DAO-QC0	DAO-QC5	DAO-QC10	DAO-QC25
GA	⁄	−47.139▽	−27.653▽	−12.505▽	3.016▲
DAO-QC0		⁄	19.486▲	34.634▲	50.155▲
DAO-QC5			⁄	15.148▲	30.669▲
DAO-QC10				⁄	15.521▲
DAO-QC25					⁄

Table 4: Protein *1b3a* average diversity delta

	GA	DAO-QC0	DAO-QC5	DAO-QC10	DAO-QC25
GA	⁄	−0.0595▽	0.0254▲	0.0569▲	0.0864▲
DAO-QC0		⁄	0.0848▲	0.116▲	0.146▲
DAO-QC5			⁄	0.0315▲	0.061▲
DAO-QC10				⁄	0.0295▲
DAO-QC25					⁄

Table 5: NK Model2 average fitness delta

	GA	DAO-QC0	DAO-QC5	DAO-QC10	DAO-QC25
GA	⁄	−57.722▽	−30.545▽	−12.078▽	1.905 -
DAO-QC0		⁄	27.177▲	45.644▲	59.627▲
DAO-QC5			⁄	18.467▲	32.450▲
DAO-QC10				⁄	13.983▲
DAO-QC25					⁄

Table 6: NK Model2 average diversity delta

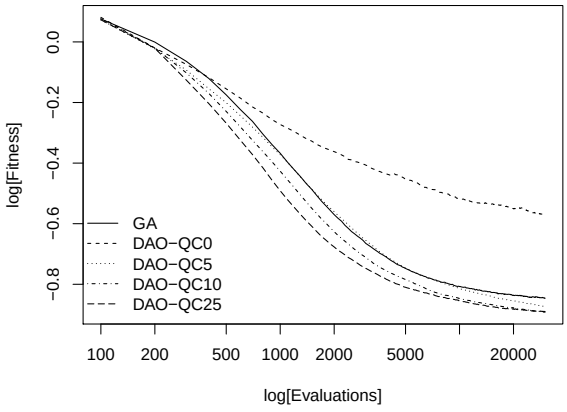


Figure 6: Convergence of average fitness of protein *1b3a*.

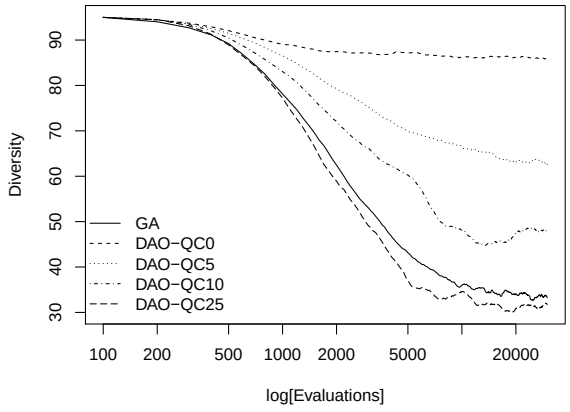


Figure 7: Convergence of average diversity of protein *1b3a*.

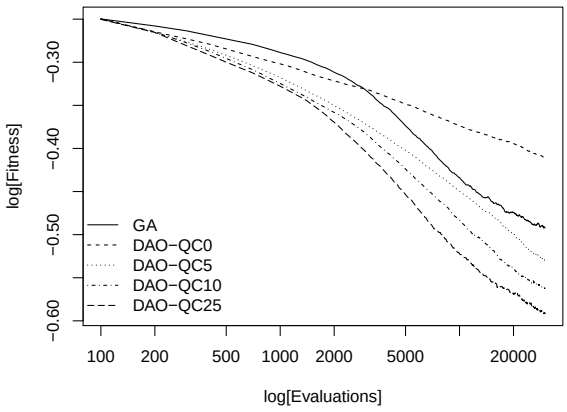


Figure 8: Convergence of average fitness of *Model 2*.

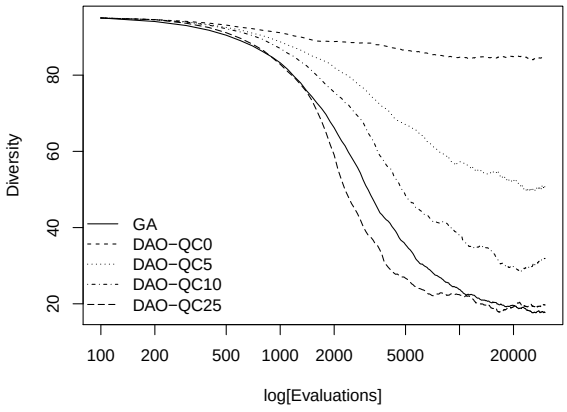


Figure 9: Convergence of average diversity of *Model 2*.

5. CONCLUSION

This article introduced a novel benchmark problem based on the well know *NK Model* extended to a *nominal discrete NKL Model* definition in a previous third-party work. Setting $L = 20$ allows the model to work with amino-acid like sequences similar to RNA with the ultimate goal of mimicking the Inverse Folding Problem (IFP). Thorough problem analysis was conducted through adaptive- and random walks in terms features like fitter neighbors, auto-correlation among others as well as an extended epistatic linkage sampling around local optima. Very similar characteristics within an upper bound of a factor two were achieved in almost all tests when comparing the *NKL Model* instances to the IFP. Running selected Genetic Algorithms with different diversity maintaining features also show very similar convergence behavior in diversity and fitness for the proposed models and the IFP. Furthermore the statistical nature of the *NK Model* with existing proofs and lemmas may provide the ground for a theoretical estimate on the number of protein sequences which fold into a given protein structure.

Acknowledgments.

Work funded by the National Research Fund of Luxembourg (FNR) as part of the EVOPERF project at the University of Luxembourg with the AFR contract no. 1356145. Experiments were carried out using the HPC facility of the University of Luxembourg [6]

6. REFERENCES

[1] S. A. Kauffman. *The origins of order: Self-organization and selection in evolution*. Oxford university press, 1993.

[2] S. A. Kauffman and E. D. Weinberger. The nk model of rugged fitness landscapes and its application to maturation of the immune response. *Journal of theoretical biology*, 141(2):211–245, 1989.

[3] R. Li, M. T. Emmerich, J. Eggermont, E. G. Bovenkamp, T. Bäck, J. Dijkstra, and J. H. Reiber. Mixed-integer nk landscapes. In *Parallel Problem Solving from Nature-PPSN IX*, pages 42–51. Springer, 2006.

[4] S. S. Nielsen, G. Danoy, W. Jurkowski, J. L. J. Laredo, R. Schneider, E.-G. Talbi, and P. Bouvry. A novel multi-objectivisation approach for optimising the protein inverse folding problem. In *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*, page (To appear). Springer, 2015.

[5] B. Rost and C. Sander. Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins*, 19(1):55–72, May 1994.

[6] S. Varrette, P. Bouvry, H. Cartiaux, and F. Georgatos. Management of an academic hpc cluster: The ul experience. In *Proc. of the 2014 Intl. Conf. on High Performance Computing & Simulation (HPCS 2014)*, Bologna, Italy, July 2014. IEEE.

[7] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.