A Novel Multi-objectivisation Approach for Optimising the Protein Inverse Folding Problem

Sune S. Nielsen[1](✉), Grégoire Danoy[1], Wiktor Jurkowski[3], Juan Luis Jiménez Laredo[4], Reinhard Schneider[2], El-Ghazali Talbi[5], and Pascal Bouvry[1]

[1] FSTC, University of Luxembourg, Walferdange, Luxembourg
{sune.nielsen,gregoire.danoy,pascal.bouvry}@uni.lu
[2] LCSB, University of Luxembourg, Walferdange, Luxembourg
reinhard.schneider@uni.lu
[3] TGAC, Norwich Research Park, Norwich, UK
wiktor.jurkowski@tgac.ac.uk
[4] LITIS, Université du Havre, Le Havre, France
jimenezj@univ-lehavre.fr
[5] INRIA Lille, Nord Europe Research Centre, Lille, France
el-ghazali.talbi@inria.fr

Abstract. In biology, the subject of protein structure prediction is of continued interest, not only to chart the molecular map of the living cell, but also to design proteins of new functions. The Inverse Folding Problem (IFP) is in itself an important research problem, but also at the heart of most rational protein design approaches. In brief, the IFP consists in finding sequences that will fold into a given structure, rather than determining the structure for a given sequence - as in conventional structure prediction. In this work we present a Multi Objective Genetic Algorithm (MOGA) using the diversity-as-objective (DAO) variant of multi-objectivisation, to optimise secondary structure similarity and sequence diversity at the same time, hence pushing the search farther into wide-spread areas of the sequence solution-space. To control the high diversity generated by the DAO approach, we add a novel Quantile Constraint (QC) mechanism to discard an adjustable worst quantile of the population. This DAO-QC approach can efficiently emphasise exploitation rather than exploration to a selectable degree achieving a trade-off producing both better and more diverse sequences than the standard Genetic Algorithm (GA). To validate the final results, a subset of the best sequences was selected for tertiary structure prediction. The super-positioning with the original protein structure demonstrated that meaningful sequences are generated underlining the potential of this work.

Keywords: Inverse Folding Problem · Protein design · Genetic Algorithm · Multi-objectivisation

A.M. Mora and G. Squillero (Eds.): EvoApplications 2015, LNCS 9028, pp. 14–25, 2015.
DOI: 10.1007/978-3-319-16549-3_2

1 Introduction

Protein engineering in general aims at designing molecules with desired properties and a method that allows to successfully design such molecules would find applications in a number of areas. For example, it could allow to design improved enzymes for biotechnology applications (e.g., waste-water treatment or biomass production), or new antibodies more specific towards already known targets (e.g., antibodies targeting a given pathogen like HIV, by binding to its envelope spikes to neutralize the virus [11]).

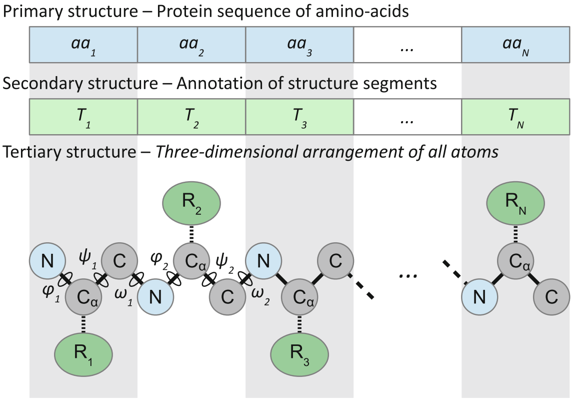


Fig. 1. Three levels of protein structure

The structure of a protein is typically represented by different levels of structures (see Fig. 1). The primary structure is the protein sequence of N amino acids (also referred to as residues) $\{aa_i\}$ where $1 \leq i \leq N$ is the residue position. The secondary structure defines the organisation of *helices*, *sheets*, *turns* and *coils* of the tertiary structure and can be expressed by a type $\{T_i\} \in \{H, S, T, C\}$ for each position i in the protein. The tertiary structure completely describes the arrangement of all atoms in the three-dimensional space. A simplified example is presented in Fig. 1 with only N and C atoms and R_i residue side-chains.

With this hierarchical definition in mind, the Inverse Folding Problem (IFP), first mentioned by Pabo in [16] can be defined as follows: given a primary structure (protein sequences) and its corresponding tertiary structure, find alternative primary structures that will result in the same tertiary structure. This makes the solution of the IFP a key part of any protein design-process, where a specific tertiary structure is targeted while keeping a certain degree of freedom in the choice of protein sequence. Furthermore, the IFP is of general scientific interest to study the size, shape and characteristics of the sequence space that matches a given target structure, and how far from the original sequence solutions can be found. In this work, the fact that matching secondary structures is a necessary, but not a sufficient condition for proteins to have the same tertiary structures is exploited to reduce the IFP to its simplest formulation: given a protein's secondary structure and its corresponding protein sequence, find a set of highly dis-similar protein sequences that could result in the most similar secondary structure. With a fast estimate of the sequence's secondary structure as objective function, the computational time can be dramatically diminished, allowing a larger part of the feasible sequence space to be explored than existing exact methods.

The resulting optimisation problem is highly *multi-modal*. Therefore the algorithm proposed in this work addresses this aspect by using a diversity measure as objective through multi-objectivisation. Additionally, the algorithm incorporates a novel constraint method that allows controlling of the high diversity induced by the multi-objectivisation approach.

The remainder of this article is organised as follows. First the current work is situated in related literature in Sect. 2, then a detailed description of the problem and the biological background is introduced in Sect. 3. In Sect. 4 the contributions of this work in terms of modeling the IFP as an optimisation problem and achieving an adjustable level of diversity in the genetic algorithm are presented. Sections 5.1 and 5.2 describe the experiments conducted and the results obtained with a validation study in Sect. 5.3. Finally the contribution, results and perspectives are summarised in Sect. 6.

2 Related Work

This section reviews some of the most relevant works related to the two main areas covered in this paper: protein design and diversity preservation in metaheuristics.

2.1 Protein Design

Since the first design of a peptide by Gutte *et al.* [8] using secondary structure rules, numerous works have described different approaches to the IFP problem. Ponder and Richards [17] used a systematic exhaustive approach of enumerating a selected subset of residue positions while Bowie *et al.* [2] introduce a 3D to 1D score at each residue position in the protein sequence. The first reported use of a Genetic Algorithm (GA) for sequence design is by Jones [9] where simplified energy and amino-acid composition terms are optimised. Until the present day the leading methods are largely based on branch-and-bound algorithms or Monte Carlo enumeration techniques, changing a limited number of residue positions [12,15,21]. Common for these methods is that they rely on evolutionary information of existing structures and use energy potential and atomic scale force-field approximations to different degrees of detail. In some works the flexibility inherent in the tertiary structure of proteins is taken into account, referred to with terms such as rotamer conformations and backbone flexibility. The complexity and exhaustive nature of most methods effectively limits the size of the sequence or decision space that can be sampled, and the final output consists of a single or few sequences close to the original sequence.

2.2 Multi-modal Optimisation and Niching

In metaheuristics, the subject of exploration vs. exploitation characteristics has been thoroughly studied. For population based optimisation algorithms it is well-known that a higher level of population diversity results in more exploration at the expense of exploitation. An elevated population diversity is especially desirable for *multi-modal*, *deceptive* and/or *dynamic* problems. In general, if diversity tends towards zero it indicates that the algorithm has converged towards a single solution, which might be an undesired behavior if it occurs too early. A number of works have sought to maintain and control diversity, e.g. crowding methods by DeJong [3], fitness sharing by Goldberg and Richardson [7], cellular algorithms by Alba and Dorronsoro [1], diversity preserving selection strategies based on hamming distance Shimodaira [20] and on altruism by Laredo *et al.* [13]. Another approach consists in designing new objectives through multi-objectivisation and thereby extending the problem to a bi- or multi-objective one. Extending problems with an objective designed specifically for diversity preservation has been proposed by Toffolo and Benini [22], Wessing *et al.* [24], as well as Deb *et al.* [5]. In these works, objectives have been designed based on the hamming distance to the closest neighbor, the distance to the nearest better and the number of individuals in the neighborhood. In this work, the diversity preserving objective is based on the average distance of each individual to all others which directly targets the global diversity measure stated by the problem, contrary to the pairwise local view of existing works. Given the discrete nature, complexity and multi-modality of the problem, an effective diversity limiting mechanism is required. The proposed approach achieves this with the added value of making the population diversity highly variable depending on a single algorithm setting.

3 Problem Description

With the focus on finding diverse solutions to the Inverse Folding Problem (IFP), we tackle a simplified model developed to matching solely the reference secondary structure - a requirement for the tertiary structure. A single solution is represented as a sequence $A = \{aa_i\}$ to consist of N residue positions, where $1 \leq i \leq N$ and $aa_i \in \{1, 2, ...20\}$ corresponds to the set of 20 possible amino-acids. As the solution space consists of a total of 20^N different combinations, considering that N is around 50–200 for typical design targets, it is clear that alternatives to exhaustive exploration are required.

3.1 Secondary Structure Estimation

The primary goal of this estimate is to obtain sequences that match the reference secondary structure. Secondary structure refers to the annotation of segmentation of the sequence into structural components, here only *helices* and *sheets* are considered. These segments are the result of the protein naturally folding so that different parts of the 3D structure connect through bonds between amino-acids on separated residue positions in the sequence. *Helices* are characterised by a corkscrew shape, *sheets* are parallel connected segments, and *loops* are everything else. Using the tool PROFphd, updated to ReProf [18], the likely secondary structure type $T_{pred}(i)$ can be predicted per amino acid aa_i in A with a reliability, $R_{pred}(i) \in \{1...10\}$ 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:

$$F_{sec}(A) = \frac{\Sigma_{max} - \sum_{i=1}^{N} M_i}{\Sigma_{max}}, \tag{1}$$

where

$$M_i = \begin{cases} 0 & \text{if } T_{pred}(i), T_{ref}(i) \notin \{helix, sheet\} \\ R_{pred}(i) & \text{if } T_{pred}(i) = T_{ref}(i) \\ -R_{pred}(i) & \text{if } T_{pred}(i) \neq T_{ref}(i) \end{cases}$$

and

$$\Sigma_{max} = \max R_{pred} \cdot |\{i | T_{ref}(i) \in \{H, E\}\}|$$

The reference types $T_{ref}(i)$ are extracted from the reference structure S_{ref}, per residue position i using the standard ‘Define Secondary Structure of Proteins’ (DSSP) algorithm [10]. As seen from Eq. 1, the calculation is only concerned with *helix* and *sheet* structures. A position i only contributes to the score if one of these are found at either $T_{ref}(i)$ or $T_{pred}(i)$ and the contribution magnitude is equal to the reliability of the prediction $R_{pred}(i)$ where match or mis-match determines the sign.

3.2 Diversity Measure

As a requirement stated in the problem description, the algorithm should not only find a single very good solution, but rather a number of good solutions as different as possible. An effective and simple measure of distance between two sequences is the Hamming-distance, defined as the number of permutations necessary to convert one into the other. Not taking gaps or varying sequence lengths into account, for two sequences $A = \{aa_i\}$ and $A' = \{aa'_i\}$ where $1 \leq i \leq N$, the Hamming distance between them is defined as:

$$d_{Hamm}(A, A') = \sum_{i=1}^{N} d_i, \; d_i = \begin{cases} 0 \text{ if } aa_i = aa'_i \\ 1 \text{ otherwise} \end{cases} . \tag{2}$$

Equation 2 states that $d_{Hamm}(A, A')$ is essentially the amount of positions one needs to change to transform A into A'. To obtain a non-negative objective value for minimisation, the average Hamming distance to all other $M - 1$ individuals in the current population, minus the sequence length N is computed:

$$F_{div}(A) = N - \frac{1}{M-1} \sum_{i=1}^{M-1} d_{Hamm}(A, A_i). \tag{3}$$

4 Methodology

With the two functions, $F_{sec}(A)$ and $F_{div}(A)$ defined for integer encoded solutions $A = \{aa_i\}$, a novel multi-objective GA based on NSGA-II [4] was applied.

Algorithm 1. DAO-QC NSGA-II

```
1: $Initialise(P_0)$                                  // randomly generated individuals
2: $t \leftarrow 0$
3: while $t < t_{max}$ do
4:    $Q_t \leftarrow makeNewPop(P_t)$                // selection, mutation, re-combination
5:    $R_t \leftarrow P_t \cup Q_t$
6:    $mutateDoubles(R_t)$                            // eliminate doubles by mutation
7:    $F \leftarrow fastNonDominatedSort(R_t)$
8:    $P_t \leftarrow truncate(F)$                    // based on domination and crowding
9:    $setQuantileConstraint(P_t)$                    // to penalise worst quantile
10: end while
```

To achieve better performance, two modifications were done which are discussed in the following and highlighted in Algorithm 1.

In the context of diversity preservation it is clear that having two or more identical individuals in the population is undesired. Hence doubles are removed by mutating them with a probability of $\frac{5}{chrom\_length}$ in Step 6, rather than eliminating them. A consequence of the nature of the objectives $F_{sec}(A)$ and $F_{div}(A)$ is that the latter is much easier to optimise, hence the population quickly consists of very diversified individuals with poor fitness according to $F_{sec}(A)$. To counter this effect the Quantile Constraint (QC) is introduced in Step 9 which prevents a precisely defined worst quantile of the population from being selected during the next generation starting in Step 4. The selection pressure can then be selectively adjusted by changing the size of the quantile C_q, which has been tested using $C_q \in \{0\,\%, 5\,\%, 25\,\%\}$.

5 Experiments

To study the performance of the proposed modified NSGA-II, with diversity objective and quantile constraint, a number of experiments have been conducted. For baseline comparison, the performance results have been compared to a standard Genetic Algorithm (GA). As final validation of the results is only really possible in the lab, an altsernative is running a top ranked protein structure prediction framework, like I-TASSER [19], on selected sequences.

5.1 Experimental Setup

Table 1 summarises the settings of the standard generational GA and the GA extended by multi-objectivisation into a Multi Objective Genetic Algorithm (MOGA) with DAO and QC. The DAO-QC MOGA version applies NSGA-II with standard selection and crossover operators: Binary tournament selection, 1-point crossover and uniform mutation. For the single objective version, a standard GA was applied using the same crossover and mutation operators. The total number of function evaluations is limited to 20000 and every experiment was repeated 30 times. As target samples, two proteins of different structural classes were chosen as reference: 256b (E. Coli Cytochrome B562) and 1b3a

Table 1. Algorithm settings

Setting	Value
Population size	100
Algorithm	NSGA-II and std GA
Termination condition	20000 function evaluations
Selection	Binary tournament (BT)
Crossover operator	1-point, $p_c = 1.0$
Mutation operator	Uniform, $p_m = \frac{1}{N}$
Quantile constraint	$C_q \in \{0\,\%, 5\,\%, 25\,\%\}$

(human C-C motif chemokine 5, RANTES). *256b* consists of $N = 106$ amino-acids packed into 4 main helices whereas *1b3a* consists of $N = 67$ amino-acids packed into 1 helix, and 3 beta-sheets as well as a long unstructured coil region.

5.2 Algorithm Results

In the following we present and compare the results observed in terms of population fitness and diversity averaged over the 30 individual runs. Figure 2 shows the convergence of the population average fitness and population diversity. Table 2(a) and (b) present the average final fitness values in numbers by pair-wise cross-comparing the three C_q settings with the GA. The Wilcoxon test indicator [25] with a 5 % significance level provides statistical confidence in comparing the sets with symbols '▲', '▽' and '-' indicating superior, inferior and no difference. The symbols refer to the column value, which is the second in each cell.

Table 2. Summary of final fitness averages

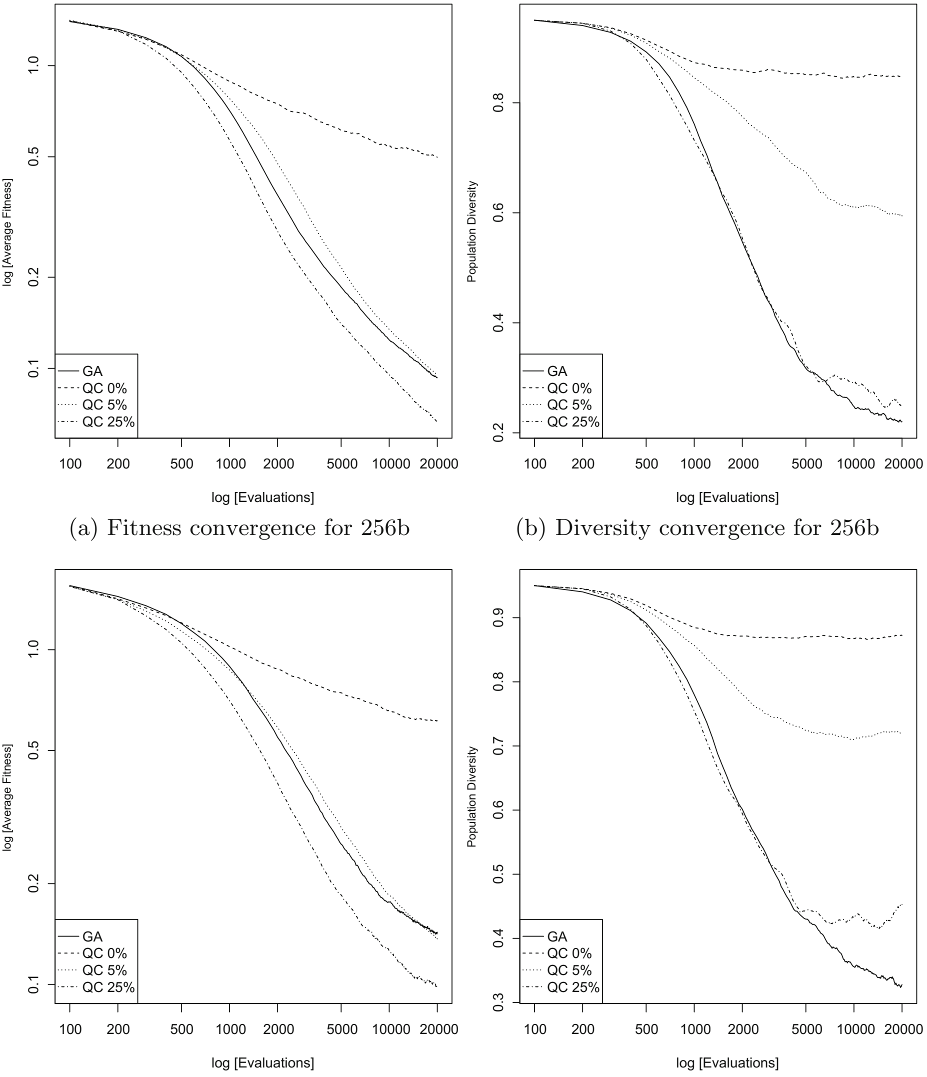
(a) 256b

	QC 0%	QC 5%	QC 25%	GA
QC 0%	/	0.498 - 0.095 ▲	0.498 - 0.066 ▲	0.498 - 0.093 ▲
QC 5%		/	0.095 - 0.066 ▲	0.095 - 0.093 ▲
QC 25%			/	0.066 - 0.093 ▽
GA				/

(b) 1b3a

	QC 0%	QC 5%	QC 25%	GA
QC 0%	/	0.613 - 0.136 ▲	0.613 - 0.098 ▲	0.613 - 0.143 ▲
QC 5%		/	0.136 - 0.098 ▲	0.136 - 0.143 ▽
QC 25%			/	0.098 - 0.143 ▽
GA				/

Clearly, the higher diversity comes at the expense of lower average fitness due to the exploration/exploitation trade-off. However the average fitness plots show that the DAO approach has better final characteristics: The $C_q = 5\,\%$ quantile



(a) Fitness convergence for 256b

(b) Diversity convergence for 256b

(c) Fitness convergence for 1b3a

(d) Diversity convergence for 1b3a

Fig. 2. Algorithm convergence

DAO-QC NSGA-II overtakes the standard GA for the *1b3a* sample scoring a significantly better final average of 0.136 vs. 0.143 with statistical confidence. For both samples with $C_q \in \{5\,\%, 25\,\%\}$ the final slope is steeper than the GA, indicating better performance given enough evaluation budget. With $C_q = 25\,\%$ the algorithm clearly outperforms the GA with statistical confidence for both samples with values 0.066 vs. 0.093 and 0.098 vs. 0.143 respectively. The steeper final slopes can be partially explained by the constantly high diversity seen in

Fig. 2. From the figure it is also evident that the size of the quantile has a direct impact on the population diversity, providing an effective tool to achieve the diversity preferred.

5.3 Validation Results

Ten generated sequences have been randomly selected between different runs per each sample (*256b* and *1b3a*) for prediction by I-TASSER [19] with the goal of assessing the meaningfulness of the generated sequences. Table 3(a) and (b) summarise super-positioning results of the predicted structures. The first column contains the standard sequence identity score [14] based on alignment with gaps. The remaining columns are well-known quality assessment metrics computed with the tertiary structure alignment tool LGA detailed in [26] with default Global Distance Test (GDT) and Longest Continuous Segments (LCS) analysis settings. The second column contains the length of the longest continuous segment N' that can be fitted below a 5A threshold after super-positioning the two structures. With $S^a = \{s_1^a, s_2^a, ...s_N^a\}$ and $S^b = \{s_1^b, s_2^b, ...s_N^b\}$ denoting the 3D positions of every residue in the two structures to compare, root-mean-square deviation (RMSD) is defined in Eq. 4, assuming the structures are optimally aligned.

$$RMSD(S^a, S^b) = \sqrt{\frac{1}{N'}\sum_i |s_i^a - s_i^b|^2}, \; i \in \{i| \; |s_i^a - s_i^b| < 5A\} \qquad (4)$$

The Global Distance Test (GDT) Total Score (TS) is a measure indicating the total average of the average percentage of residues that can be fitted below each of the thresholds $\{0.5A, 1.0A, 1.5A, ...10.0A\}$. The final column is a quality estimate where values below 2.0 indicates a rather weak alignment (for further details please refer to [26]). Figure 3(a) and (b) show a graphical super-positioning of the best scoring generated sequences of Table 3 with PyMol [6]. Overall the sequences generated for the first sample *256b* do better than the *1b3a* sample, which can also be seen visually. This can largely be explained by the fact that only secondary structure prediction has been used, and that the

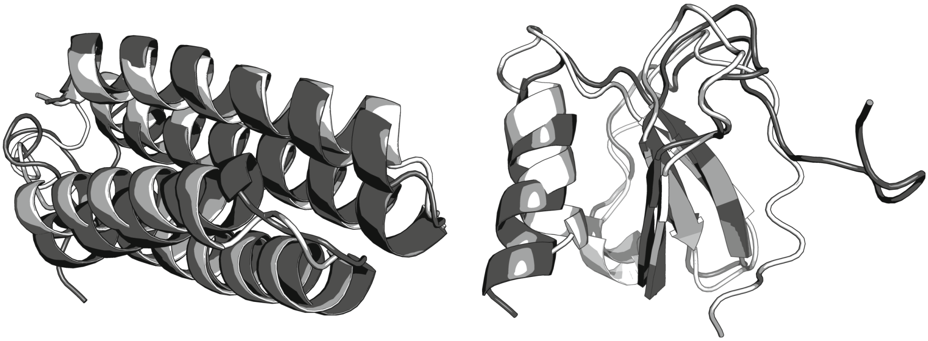
Table 3. Summary of alignment scores of selected sequences

(a) 256b

ID[%]	$N' < 5$A	RMSD	GDT_TS	LGA_Q
8.65	92	2.82	60.142	3.148
12.38	98	2.19	69.811	4.276
11.43	106	3.09	61.557	3.325
14.42	104	2.21	75.943	4.506
8.57	53	3.53	37.264	1.458
10.48	94	3.36	55.425	2.72
17.14	104	2.65	70.283	3.776
10.48	39	2.62	32.783	1.432
14.42	47	2.15	38.208	2.089
11.54	64	3.08	46.934	2.014

(b) 1b3a

ID[%]	$N' < 5$A	RMSD	GDT_TS	LGA_Q
10.94	36	3.04	39.552	1.148
6.35	47	3.18	54.104	1.432
10.94	36	3.22	40.299	1.083
16.92	41	3.71	44.776	1.076
12.70	37	3.1	38.806	1.157
7.94	49	2.96	54.478	1.603
9.23	34	3.12	35.821	1.055
14.06	56	2.63	67.91	2.05
15.00	32	3.39	38.433	0.916
12.70	49	2.94	51.493	1.614


(a) 256b

(b) 1b3a

Fig. 3. Predicted (lighter) and reference (darker) structures superpositioned

latter sample contains less structured elements and mainly beta-sheet segments. The differences are mostly in coil regions and due to slight mispositioning of the main structure elements. For the highly structured *256b*, the best prediction gave a global score (GDT_TS) of almost 76 % with most residues within 5A. The non-structured coil of *1b3a* diverts away from the target, giving a best GDT_TS of less than 68 %, but a low RMSD of 2.63 of the 56 residues that fit below 5A. Overall about half of the generated sequences were predicted with a total score above 50 % and an RMSD below or close to 3A, which is reasonable considering that only secondary structure was optimised and that the main part have sequence identity below 15 % with a minimum of 6.35 % - much lower than existing approaches where values below 25 % are rare.

6 Conclusion

In this paper we have presented a novel approach to find a large amount of protein sequences that may result in a given reference 3D structure. This problem, referred to as the Inverse Folding Problem (IFP), has received a lot of attention in theoretical chemistry and biophysics over the last 30 years, mostly for its potential application in protein design. It is also of interest to study the extent of the sequence space that may produce similar tertiary structures, and how far from the original reference sequence such solutions can be found.

By defining the task as finding highly diverse sequences with most similar secondary structures, an optimisation problem was modeled to find many well-scoring sequences in a few hours, which is fast compared to state-of-the-art methods. To achieve high diversity we have adapted the requirement as an additional objective and extended the problem through *multi-objectivisation* to become Multi-Objective with Diversity-As-Objective (DAO). Combining the novel Quantile Constraint (QC) with the DAO approach allows to shift focus arbitrarily between diversity or fitness, and final results found are comparable or better than the standard GA on average, while the diversity of found sequences remains higher at the same time. In addition, the algorithm convergence was observed as being steeper than the standard GA which promises very good solutions given an evaluation budget beyond the computational limitations set in this work.

Selected sequences of the highly diverse sets generated were inspected further by predicting their structure with I-TASSER (a top ranked structure prediction software). Final validation was done by comparing the predicted structures to their respective reference by tertiary structure super-position. For both samples *256b* and *1b3a* meaningful predictions were generated with close to 76 % and 68 % GDT_TS scores respectively and RMSD well below 3A. As could be expected, the method works better for the sample with more defined secondary structure, and less well in coil regions which are not captured by the objective function.

Future and ongoing works will address the identification of those sequences that actually fold into the reference structure by designing new objectives and constraints also addressing coil regions. Independent of this, sequences found could already be used as starting points for other exact protein design methods and possibly generate successful designs with a very low sequence identity comparing to the reference.

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 [23]

References

1. Alba, E., Dorronsoro, B.: The exploration/exploitation tradeoff in dynamic cellular genetic algorithms. IEEE Trans. Evol. Comput. **9**(2), 126–142 (2005)
2. Bowie, J.U., Lüthy, R., Eisenberg, D.: A method to identify protein sequences that fold into a known three-dimensional structure. Science (New York, N.Y.) **253**(5016), 164–170 (1991)
3. De Jong, K.A.: Analysis of the behavior of a class of genetic adaptive systems. Ph.D. thesis, University of Michigan Ann Arbor, MI, USA (1975)
4. Deb, K., Agrawal, S., Pratap, A., Meyarivan, T.: A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II. Lect. Notes Comput. Sci. **849–858**, 2000 (1917)
5. Deb, K., Saha, A.: Finding multiple solutions for multimodal optimization problems using a multi-objective evolutionary approach. In: Proceedings of the 12th Annual Conference on Genetic and Evolutionary Computation, pp. 447–454. ACM (2010)
6. DeLano, W.L.: The pymol molecular graphics system, delano scientific, San Carlos, CA, USA (2002). There is no corresponding record for this reference (2002)
7. Goldberg, D.E., Richardson, J.: Genetic algorithms with sharing for multimodal function optimization. In: Genetic Algorithms and Their Applications: Proceedings of the Second International Conference on Genetic Algorithms, pp. 41–49 (1987)
8. Gutte, B., Däumigen, M., Wittschieber, E.A.: Design, synthesis and characterisation of a 34-residue polypeptide that interacts with nucleic acids. Nature **281**(5733), 650–655 (1979)

9. Jones, D.T.: De novo protein design using pairwise potentials and a genetic algorithm. Protein Sci. **3**, 567–574 (1994)
10. Kabsch, W., Sander, C.: Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers **22**(12), 2577–2637 (1983)
11. Klein, F., Mouquet, H., Dosenovic, P., Scheid, J.F., Scharf, L., Nussenzweig, M.C.: Antibodies in HIV-1 vaccine development and therapy. Science (New York, N.Y.) **341**(6151), 1199–1204 (2013)
12. Klepeis, J.L., Floudas, C.A., Morikis, D., Tsokos, C.G., Lambris, J.D.: Design of peptide analogues with improved activity using a novel de novo protein design approach. Ind. Eng. Chem. Res. **43**(14), 3817–3826 (2004)
13. Jiménez Laredo, J.L., Nielsen, S.S., Danoy, G., Bouvry, P., Fernandes, C.M.: Cooperative selection: improving tournament selection via altruism. In: Blum, C., Ochoa, G. (eds.) EvoCOP 2014. LNCS, vol. 8600, pp. 85–96. Springer, Heidelberg (2014)
14. Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., Higgins, D.G.: Clustal w and clustal x version 2.0. Bioinformatics **23**(21), 2947–2948 (2007)
15. Mitra, P., Shultis, D., Brender, J.R., Czajka, J., Marsh, D., Gray, F., Cierpicki, T., Zhang, Y.: An evolution-based approach to de novo protein design and case study on Mycobacterium tuberculosis. PLoS Comput. Biol. **9**(10), e1003298 (2013)
16. Pabo, C.: Molecular technology: designing proteins and peptides. Nature **301**(5897), 200 (1983)
17. Ponder, J.W., Richards, F.M.: Tertiary templates for proteins: use of packing criteria in the enumeration of allowed sequences for different structural classes. J. Mol. Biol. **193**(4), 775–791 (1987)
18. Rost, B., Sander, C.: Combining evolutionary information and neural networks to predict protein secondary structure. Proteins **19**(1), 55–72 (1994)
19. Roy, A., Kucukural, A., Zhang, Y.: I-TASSER: a unified platform for automated protein structure and function prediction. Nat. Protoc. **5**(4), 725–738 (2010)
20. Shimodaira, H.: Dcga: a diversity control oriented genetic algorithm. In: ICTAI, pp. 367–374 (1997)
21. Smadbeck, J., Peterson, M.B., Khoury, G.A., Taylor, M.S., Floudas, C.A.: Protein wisdom: a workbench for in silico de novo design of biomolecules. J. Vis. Exp. **77**, e50476 (2013)
22. Toffolo, A., Benini, E.: Genetic diversity as an objective in multi-objective evolutionary algorithms. Evol. Comput. **11**(2), 151–167 (2003)
23. Varrette, S., Bouvry, P., Cartiaux, H., Georgatos, F.: Management of an academic HPC cluster: the UL experience. In: Proceedings of the 2014 International Conference on High Performance Computing & Simulation (HPCS 2014), Bologna, Italy. IEEE, July 2014
24. Wessing, S., Preuss, M., Rudolph, G.: Niching by multiobjectivization with neighbor information: trade-offs and benefits. In: 2013 IEEE Congress on Evolutionary Computation (CEC), pp. 103–110. IEEE (2013)
25. Wilcoxon, F.: Individual comparisons by ranking methods. Biometrics Bull. **1**(6), 80–83 (1945)
26. Zemla, A.: LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Res. **31**(13), 3370–3374 (2003)