Protein Design by Multiobjective Optimization: Evolutionary and Non-Evolutionary Approaches

Sandeep V. Belure

Center for Advanced Biotechnology
and Medicine
Rutgers University
NJ, USA
sandy.belure@gmail.com

Ofer M. Shir

Computer Science Department,
Tel-Hai College, and
The Galilee Research Institute - Migal
Upper Galilee, Israel
ofersh@telhai.ac.il

Vikas Nanda

Center for Advanced Biotechnology
and Medicine
Rutgers University
NJ, USA
nanda@cabm.rutgers.edu

ABSTRACT

Traditional simulation-based protein design considers energy minimization of candidate conformations as a singleobjective combinatorial optimization problem. In this paper we consider a challenging protein design problem, producing twelve protein species based on collagen that uniquely assort into four groups of three: a problem defined herein as a 4-level heterotrimer. We formulate a bi-objective combinatorial minimization problem that targets both stability and specificity of the 4-level heterotrimer. In order to approximate its Pareto frontier, we utilize both evolutionary and non-evolutionary approaches, operating in either Pareto or aggregation fashions. Our practical observations suggest that the SMS-EMOA with Evolution Strategies' operators is more effective than standard heuristics deployed in computational protein design, such as Simulated Annealing, Replica Exchange or the Canonical Genetic Algorithm. We investigate the attained Pareto optimal sets using Barrier Tree analysis, aiming to provide insights into the chemical search-space, as well as to explain the observed algorithmic trends. In particular, we identify Replica Exchange as a promising non-evolutionary technique for this problem class, due to its efficient exploration capabilities. Overall, a common high-level protocol for simultaneous landscape analysis of evolutionary and non-evolutionary search methodologies is put forward for the first time.

KEYWORDS

Protein design, simulation-based multiobjective optimization, evolutionary algorithms, replica exchange, simulated annealing, combinatorial landscapes, barrier trees, negative slope coefficients

ACM Reference format:

Sandeep V. Belure, Ofer M. Shir, and Vikas Nanda. 2017. Protein Design by Multiobjective Optimization: Evolutionary and Non-Evolutionary Approaches. In *Proceedings of GECCO '17, Berlin, Germany, July 15-19, 2017*, 8 pages.
DOI: <http://dx.doi.org/10.1145/3071178.3071268>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

GECCO '17, Berlin, Germany

© 2017 ACM. 978-1-4503-4920-8/17/07...\$15.00

DOI: <http://dx.doi.org/10.1145/3071178.3071268>

1 INTRODUCTION

The construction of synthetic proteins is an emerging technology potentially impacting a number of sectors from medicine to industrial process chemistry to nanoscience [8, 14, 22]. In nature, proteins perform a wide range of biological roles acting as motors, energy producers, catalysts, sensors, structural scaffolds and so on. This tremendous functional plasticity exists despite the seemingly limited twenty amino acid alphabet which protein chains are composed from. Given state-of-the-art technologies in gene construction and protein expression – it is possible to produce proteins of nearly any desired amino acid sequence, facilitating the development of synthetic proteins tailored for non-natural applications. However, the challenge remains to specify the optimal amino acid sequence for a target molecule. Given that the typical protein is over 100 amino acids in length, there are over 20^{100} ($1.2 \cdot 10^{130}$) possible sequences to choose from. For solution spaces of such cardinality, enumeration is impossible in the laboratory or computationally, requiring effective algorithms to treat it.

Efforts to design proteins have spurred advances in algorithm and software development that address the large combinatorial challenge of choosing the appropriate sequence. Each amino acid sequence corresponds to a set of three-dimensional (3D) molecular structures. The ideal protein sequence should map to a limited number of 3D structures, which should in turn present amino acids in a spatial configuration to perform the desired function. Calculations that assess the quality of 3D atomic models are computationally intensive. As such, protein design algorithms must address both sampling in a large sequence space and efficiency of evaluating the quality of 3D structures [13]. These are integrated into design software platforms, e.g., ROSETTA [15] and protCAD [25].

Evolutionary algorithms (EAs) [1], powerful stochastic global search methods gleaned from the model of organic evolution, have been for several decades successful in treating high-dimensional optimization problems. They especially excel in scenarios where quality evaluation provided by computer-based simulation constitutes the objective function, also referred to as *simulation-based optimization*, or in black-box evaluations, such as in *experimental optimization*. In material science, EAs have been widely applied to simulation-based global cluster geometry optimization [12] at the larger scale, and to molecular design [5] at the smaller scale, the latter comprising small-molecule geometry optimization, protein-ligand docking, and protein folding, to name a few.

Here are the proposed contributions of the current study:

- (1) Multiobjective protein design formulation: stability versus specificity

- (2) Pareto algorithmic treatment: evolutionary versus non-evolutionary approaches; especially, Replica Exchange as a solver

- (3) Solutions investigation and landscape analysis

The remainder of this paper is organized as follows. In Section 2 we outline the protein design challenge while focusing on the multiobjective perspective – we specify our notation, and formulate the bi-objective protein design problem of interest. We then describe our approaches to solving the prescribed challenge in Section 3, namely single-objective heuristics for an aggregated objective function, versus selected evolutionary multiobjective optimization algorithms (EMOAs). Our practical observations on the current case-study are reported in Section 4, where we also investigate the attained solutions and conduct landscape analysis. Finally, we summarize our work and findings in Section 5, where we also draw possible directions for future work.

2 MULTIOBJECTIVE PROTEIN DESIGN

In the majority of cases, protein design protocols seek to optimize a single objective – identifying a sequence of amino acids that optimizes the stability of a target state relative to an unfolded reference state – namely, the free energy of protein folding. Stability calculations may depend on an explicit 3D atomistic model of the protein or simplified coarse grained or discrete structure-sequence maps that reduce the computational cost. Such design efforts assume that competing, off-target structural states are unlikely to be formed when the target state has a low energy, an approach termed “forward design” by Fleishman and Baker [11]. This results in designs that are both stable and specific for the target conformation.

Another approach rooted in statistical mechanics is to compute the stability of a sequence across an enumeration of conformation states representing the ensemble of all possible structures. For proteins of any reasonable size, the total number of states is too large to be explicitly computed, requiring assumptions as to which competing states are relevant to be included in the partition function. This approach has been used by a number of groups to design proteins that specifically form target oligomeric states (see, e.g., [33, 34]), or in special cases – a sequence that forms different target states depending on solution conditions [16]. Depending on the problem formulation, the optimization problem may be reduced in this approach to a single objective function, which reflects the Boltzmann probability density function of forming a target state over the total *partition function*. If, however, the heuristic approach to address competing states is not rooted in a rigorous treatment of energetics, or if other design objectives are considered in addition to stability – reduction to a single objective function then becomes much more challenging. Nivón et al. [23] explored issues with designing mutations that improved the stability of a protein without perturbing its structure relative – resulting in an objective function that combined both energetic and geometric terms. Fleishman examined using fuzzy logic operators as a strategy for integrating disparate variables into a single objective function [31]. Factors such as cell permeability, toxicity, cost-of-production and so forth are potential parameters to optimize in multiobjective protein design. Therefore, advances in multiobjective optimization within the field of protein engineering and design are much needed.

Figure 1: CMP heterotrimer design approach. (a) Interactions amongst individual peptide chains in the 3D triple-helix structure of collagen are represented through a contact map. G-x-y are the repeated triplets where G = glycine and x, y are other variable amino acids. The blue peptide strand’s sequence is represented twice in the bottom panel of the figure to show the 3D molecule’s contact map in 2D. (b) Stability, modeled by the Energy score E_T , is computed as the sum of contacts for the target heterotrimer. Specificity is modeled by the Gap score E_{gap} as the difference between target and most stable competing state energies. Based on experience with prior designs, a wide range of E_T and E_{gap} values are possible.

2.1 4-Level Heterotrimer Orthogonality

Natural collagen is a fibrous protein composed of 3 separate polypeptide chains that associate into a 3-stranded super-helix or triple-helix (Fig. 1). Each chain is composed of a repeating sequence triplet required for collagen to assemble. In natural fibrillar collagens, this triplet can repeat over three-hundred times, resulting in over one thousand residue long collagen molecules. The three chains may have identical sequences – identified as “homotrimers”. Alternatively, there are examples of collagens incorporating mixtures of two or three different sequences – “heterotrimers”. Studying how the composition of collagen is controlled is important for its biological function. Challenges associated with working with natural collagen, have motivated researchers to utilize collagen mimetic peptides (CMPs). These typically around thirty-residue long peptides synthesized in the laboratory are much shorter than natural collagen, but adopt the same 3D structure [2]. CMPs also incorporate repeating G-x-y triplet repeats. By varying choice of

amino acids at the x and y positions, it is possible to control stability and heterotrimer composition of the CMP 3D structure [33, 34]. The design of such systems is the focus of the current work.

Identifying sequences that form stable and specific heterotrimers is a significant challenge for CMP design. In forming an abc heterotrimer composed of three chains a, b and c, each with a unique sequence, one must nominally consider $3^3 = 27$ unique structural states (aaa, aab, aac, ..., ccc). The ideal choice of sequences for a, b and c will maximize the stability of abc without forming significant populations of the competing twenty-six states. This sequence design challenge has been computationally addressed using a discrete representation of the triple-helix that accounts for interactions between x and y positions on adjacent chains [10, 34]. More recently, two heterotrimers were designed that assembled independently in a mixture of six peptides {a,b,c,d,e,f} into abc and def [33]. This has motivated us to investigate what practical limits exist on the complexity of a set of interacting CMPs.

In existing heterotrimer designs, sequences are composed of ten G-x-y triplets, resulting in thirty-residue long CMPs. The choice of triplets is limited to 4 possibilities {DOG, KOG, PDG, PKG}, allowing for extending networks of electrostatic interactions between positive K (lysine) residues and negative D (aspartic acid) residues. Thus, for a 1-level abc heterotrimer, there are 4^{30} possible designs. As the number of interacting peptides increases to 6 (a 2-level heterotrimer), the sequence space becomes significantly larger (4^{60}) and the number of combined homo- and heterotrimer associations increases to 216. In order to identify sequences that both stabilize the target state and disfavor competing states, two objectives – namely, stability and specificity (Fig. 1b) – are optimized during sampling of CMP sequences.

We explore the design of a 4-level heterotrimer consisting of twelve interacting peptides intended to form abc, def, ghi and jkl. Validation of designs in the laboratory has proved challenging as the increasing problem complexity has resulted in both suboptimal stability and specificity for the peptides of interest. One of the challenges is that with four target trimer states, there are nominally eight objectives – the stability E_j and specificity G_j , per each state $j = 1 \dots 4$. All objectives constitute functions of a given sequence $\vec{d}$, which adheres to the triplets' prescribed alphabet, $\vec{d} \in \mathcal{D} = \{\text{KOG}, \text{DOG}, \text{PDG}, \text{PKG}\}^{120}$, that is $E_j = E_j(\vec{d})$, $G_j = G_j(\vec{d})$ for $j = 1 \dots 4$. To reduce the number of objectives, we take the mean stability and specificity for each target, resulting in two global measures of anticipated system performance, to be formulated shortly. As mentioned before, each one of the twelve peptide strands is composed of ten amino acid triplets, representing a thirty-residue long monomer chain. Thus, every feasible candidate solution is now represented by 120 triplets and is associated to two objective values. For calculating the objective means, *geometric mean* of the individual objectives was preferred over *arithmetic mean* – this serves to reduce the variance among the four target state energies and gap values being calculated. Our bi-objective protein design problem can be formulated as:

$$\begin{aligned} &\text{minimize}_{\vec{d} \in \mathcal{D}} & f_1(\vec{d}) &= (E_1 \cdot E_2 \cdot E_3 \cdot E_4)^{\frac{1}{4}} \\ &\text{minimize}_{\vec{d} \in \mathcal{D}} & f_2(\vec{d}) &= (G_1 \cdot G_2 \cdot G_3 \cdot G_4)^{\frac{1}{4}} \\ &\text{subject to:} & & \{\text{holonomic constraints}\}, \\ & & & \mathcal{D} = \{\text{KOG}, \text{DOG}, \text{PDG}, \text{PKG}\}^{120}. \end{aligned} \quad (1)$$

Stability parameter calculation for a target state trimer was similar to that of a trimer state's stability calculation elaborated in [34]. The target state was constrained to have a net-neutral charge by adding a charge-proportionate *penalty term* to the target state stability, but not to competing state calculations, during the sequence search process. Specificity (i.e., gap) calculations were computed as the difference between stabilities of the target state and the best energy competing state. Competing states were defined as any trimer association state other than the target state that had one or more of the target trimer state's three peptide chains – among the $12 \times 12 \times 12 = 1728$ possible 3 chain stoichiometric associations.

Previous designs on a 1-level heterotrimer were developed by Simulated Annealing (SA) [13, 34] or Genetic Algorithms (GAs) [10] approaches. The current work compares these two approaches along with Replica Exchange (RE) [9, 20], which has been used in singleobjective sequence design [17]. We compare these to EMOAs on the more challenging target of designing a 4-level heterotrimer.

3 OPTIMIZATION SOLVERS

3.1 Singleobjective Heuristics

Despite its known drawbacks [6], aggregating a multiobjective problem into a singleobjective consideration is routinely exercised within the biochemistry community and therefore discussed herein. To this end, a single cost function C is defined as the weighted aggregate of the 2 objectives under consideration:

$$\text{minimize}_{\vec{d} \in \mathcal{D}} C(\vec{d}) = w_1 \cdot f_1(\vec{d}) + w_2 \cdot f_2(\vec{d}). \quad (2)$$

Preliminary runs were carried out to calibrate the scaling parameters for normalization: $\{w_1, w_2\} = \{9 - \alpha, \alpha\}$, $\alpha = 0, 1, \dots, 9$. The search algorithms used for this approach were the Canonical GA, SA and RE – as outlined below.

3.1.1 Canonical Genetic Algorithm - GA. For GA, an input parent sequence set for crossover and mutation was represented by a sequence block of 12 rows and 10 triplet columns (a total of 120 triplets), with each row representing a monomer peptide sequence and every three non-overlapping sets of consecutive rows representing a target state trimer sequence (the first three rows contained the sequences of the 1st target state trimer and so on). Crossover points were selected in alternate simulation cycles at the mid-points of the sequence block row height (retaining individual parent target state trimer stability arising from unperturbed individual target state trimer contact map) and the sequence block column width (retaining parent trimers either left axial half or right axial half axial specificity arising from preserved left or right half trimers' contact map). The number of parent sequence sets selected for crossover in each generation of GA was 2.

The sequence block was reset every *epoch* to a randomly generated sequence set to promote diversity within the gene pool.

3.1.2 Simulated Annealing - SA. SA consisted of a standard protocol of a single temperature parameter T , which undergoes cooling over the course of several sequence triplet mutation and corresponding evaluations. Selection followed the Metropolis-Hasting criterion, i.e., cost deteriorations were accepted with a probability of $\exp(-\Delta C/T)$, with ΔC denoting the cost difference between the offspring and parent.

3.1.3 Replica Exchange - RE. RE, also known as Parallel Tempering Monte-Carlo [9], is a search heuristic that addresses challenges of molecular dynamic sampling of 3D conformations of proteins [4]. It is generally applied to rough energy search landscapes as is usually the case for atomic scale energy parameter calculations. In RE, two random parent sequence sets are initialized and subjected to two constant temperature values – one high and one low – and a series of triplet mutations are carried out on the two initialized parent sequence sets. The high temperature simulation counterpart is able to hop across large energy barriers easily, as would have been the case during the initial starting phase of SA, and is thus able to explore promising regions of the landscape. The low temperature simulation counterpart is able to exploit more refined “valley” regions of the landscape, similarly to the latter phase of SA. Importantly, an *exchange* occurs every *epoch*: either the sequence configurations or the two temperatures are exchanged, based on a certain probability value. If the higher temperature simulation is able to discover an offspring sequence set with a better cost function value than the one arrived at by the lower temperature simulation counterpart, the exchange is accepted with a probability 1 (always accepted). If not, the exchange may still be carried out with a probability of $\exp\left(-\frac{\Delta C}{(1/T_{\text{low}} - 1/T_{\text{high}})}\right)$, with ΔC representing the cost difference between the high-temperature offspring and the low-temperature offspring. We expected RE to be effective as sequences generated during the high temperature exploration phase act as seed sequences for local-search during the low temperature exploitation phase.

3.2 Pareto-Based Approaches

Let us define the following multiobjective vector minimization problem: minimize $\vec{y}(\vec{d}) = \{f^{(1)}(\vec{d}), \dots, f^{(k)}(\vec{d})\}$ subject to $\vec{d} \in \mathcal{D}$, where $\mathcal{D}$ is a set of feasible solutions (within a prescribed combinatorial space). $\mathcal{Y} \subset \mathbb{R}^k$ is its image in the objective space, and $\vec{f}: \mathcal{D} \rightarrow \mathcal{Y}$ is an objective function that projects $\mathcal{D}$ onto $\mathcal{Y}$. If $\vec{d} \in \mathcal{D}$, then $\vec{y} = \vec{f}(\vec{d}) \in \mathcal{Y}$ and the value of the i^{th} objective of solution $\vec{d}$ is $y^{(i)} = f^{(i)}(\vec{d})$. For $\vec{d}_1, \vec{d}_2 \in \mathcal{D}$, $\vec{y}_1 = \vec{f}(\vec{d}_1), \vec{y}_2 = \vec{f}(\vec{d}_2) \in \mathcal{Y}$, let $\vec{y}_1 \equiv \vec{y}_2$ and $\vec{y}_1 \leq \vec{y}_2$ if $y_1^{(i)} = y_2^{(i)}$ and $y_1^{(i)} \leq y_2^{(i)}$ for all i , respectively. Two solutions $\vec{d}_1, \vec{d}_2 \in \mathcal{D}$, $\vec{y}_1 = \vec{f}(\vec{d}_1), \vec{y}_2 = \vec{f}(\vec{d}_2) \in \mathcal{Y}$ are equivalent if $\vec{y}_1 \equiv \vec{y}_2$. We write $\vec{y}_1 < \vec{y}_2$ if $y_1^{(i)} < y_2^{(i)}$ for all i and $y_1^{(i)} < y_2^{(i)}$ for some i . In this case, we would say that $\vec{y}_2$ is Pareto dominated by $\vec{y}_1$, or $\vec{y}_1$ Pareto dominates $\vec{y}_2$. The individual Pareto-ranking of a given candidate solution is defined as the number of other solutions dominating it. The crucial claim is that for any compact subset of $\mathbb{R}^k$, there exists a non-empty set of minimal elements with respect to the partial order $\leq$. The goal of Pareto optimization is thus to obtain the *non-dominated set*, the so-called Pareto frontier $\mathcal{F}$, and its pre-image within $\mathcal{D}$, the so-called *Pareto optimal set*. The computational problem of attaining the Pareto Frontier of a multiobjective optimization problem can be treated by means of computational intelligence heuristics (possibly evolutionary), or alternatively, by formal algorithms utilizing mathematical programming solvers (e.g., the so-called Diversity Maximization Approach [21]). Here, we are especially interested

in EMOAs, which have undergone considerable development in the last two decades [7]; next, we describe 2 selected techniques.

3.2.1 SPEA2-Based Approach. Strength Pareto Evolutionary Algorithm 2 (SPEA2) [35] was used. Solutions obtained from a hybridization of RE + GA + SA were archived at regular intervals using SPEA2’s non-dominated set archiving code. For crossover and mutation, a binary tournament selector with a tournament size of 2 was used to select the 2 parent solution sets from among the previous generation’s archived solutions. Sequences with good f_1 values were preferred over sequences with good f_2 values in the event of a conflict arising in archiving an individual to a fixed-size archive. We used an archive of size 20.

3.2.2 SMS-EMOA with MIES Operators. The SMS-EMOA [3] with a mixed-integer evolution strategy (MIES) [27] solver was also used. In short, SMS-EMOA initializes a random population of candidate solutions, and iteratively generates new individuals by *mutation* and *recombination*. We employ MIES-based self-adaptive mutation operator accounting to the categorical variables $\vec{d}$, $\forall d_i \in \mathcal{D}_i$, wherein each individual carries a strategy parameter $p_d \in [0, 1]$. The details concerning the mutation operator are provided in Algorithm 1. The recombination operator is applied only with a probability $p_c = 0.1$: *discrete recombination* is applied to the decision variables, whilst intermediate recombination is applied to the strategy parameter. We consider a steady-state approach, where in each iteration a single offspring is generated, after which a single individual is eliminated from the population, based upon multiobjective selection, comprising two phases: sorting the Pareto domination ranking, followed by sorting the hypervolume contribution (S-metric) per rank. Altogether, our implementation and parameter settings mostly followed [26], using a population of size 20.

```

MIESmutate( $\vec{d}, p_d, n_d$ )
 $\tau^{(d)} \leftarrow \frac{1}{\sqrt{2 \cdot n_d}}$ 
 $p'_d \leftarrow 1 / \left[ 1 + \frac{1-p_d}{p_d} \cdot \exp\{-\tau^{(d)} \cdot \mathcal{N}(0, 1)\} \right]$ 
for  $i = 1, \dots, n_d$  do
    if  $\mathcal{U}(0, 1) < p'_d$  then
         $d'_i \leftarrow$  uniformly randomly a value from  $\mathcal{D}_i$ 
    end
end
return  $\{\vec{d}', p'_d\}$ 

```

Algorithm 1: MIES-based self-adaptive mutation operator utilized by the SMS-EMOA’s solver: $\{\vec{d}, p_d\}$ are the (categorical) decision variables and the strategy parameter, respectively, where $\mathcal{D}$ is the set of available categories per variable. $\mathcal{N}$ and $\mathcal{U}$ denote the normal and uniform distributions, respectively.

4 PRACTICAL OBSERVATIONS

The five algorithmic approaches performed differently on the bi-objective use-case, as evident in Fig. 2. Among the singleobjective heuristics, RE performed best in terms of optimizing stability and

Figure 2: Performance of tested algorithms on 4-level heterotrimer design. The first 3 algorithms GA, SA, RE used an aggregated approach (Eq. 2). This simulation was run twice to show 10 steps $\times$ 2 = 20 overlapping sequence solution sets for each algorithm. The 4th and 5th algorithms constitute Pareto-based approaches. For consistency of performance comparison across algorithms, the total number of objective function evaluations was kept constant at $5 \cdot 10^6$ calls.

Figure 3: Similarities between DGs and BTs (see also [30]). a) Willow Tree: A gently sloping funnel featuring high barriers, b) Palm Tree: A steeper funnel featuring low barriers, and c) Banyan Tree: A rough energy landscape.

specificity over a fixed budget of function evaluations. RE effectively sampled minima of both individual objectives with the aid of the higher temperature near random walk sequence-space exploration. SA performed better than the GA, possibly due to its dependency upon single triplet mutations compared to that of GA's

Figure 4: The contact map for the two extreme solution points (i.e., indexed 1, 2, 10, 11). The positively-charged part of the peptide, populated by either PKG or KOG triplet is shown in green and yellow respectively, whereas the negatively-charged part, populated by either PDG or DOG is shown in pink or orange respectively. Salt bridges indicating attractive interactions (K-D) are depicted in blue and the repulsive interactions (K-K, D-D) are shown in red. The 4 target state trimers along with the most stable competing state trimer are shown.

larger sequences length crossovers. Moreover, an unfavorable vertical mid-point crossover in the GA may have also disrupted networks of favorable interactions, reducing the range of accessible objective values. A large number of Pareto-dominated solution-points were identified in the final datasets of the three algorithms.

SMS-EMOA had the best performance in terms of accuracy, i.e., converging to the Pareto optimal set of the highest quality. The coverage of the Pareto frontier obtained by SMS-EMOA was limited to a sub-range of energies and gaps when compared to the SPEA2-based hybridization. This is explained by the fact that solution sets at the extremal objective function values are either nearly dominated (Fig. 1b, and beyond -133 and upto -120 value range in Fig. 2), or are reachable in the hybridization approach by setting one of the weights $\{w_1, w_2\}$ in Eq. 2 to 0. At the same time, SMS-EMOA's observed convergence rate was slower in an order of magnitude. The absence of the S -metric criterion was evident in the hybridization approach: unlike SMS-EMOA, the solution-points were not evenly spaced on the frontier.

Finally, the extreme points of the Pareto optimal set, two per each objective function, are explicitly provided in Fig. 4.

4.1 Chemistry and Landscape Analysis

Based on energies and gaps for experimentally validated 1-level and 2-level designs, the 4-level designs produced in this analysis are promising candidates for examination in the laboratory. Stability values around -135 generally correspond to CMPs with melting temperatures around 25° - 30°C in the laboratory. The gaps are smaller than the lower-level designs and the potential for competing mixtures to form is a possibility. It may be that the conflict

between the two objectives constitutes a practical barrier to the realization of a well-behaved 4-level design. This question can only be answered experimentally. Previous 1-level design simulations, utilizing SA to minimize a similar aggregate cost function, were observed to converge rapidly (within a few hundred objective function calls), and produced CMPs that well-behaved in the laboratory [34]. In the current 4-level use-case, the number of objective function calls required to arrive at reasonable solutions is significantly larger, and performance is strongly dependent upon the algorithmic approach. To understand why this may be the case, we examined the energy landscape of solutions along the Pareto frontier, as was attained by the 4th and 5th approaches herein.

Fitness landscapes in the peptide and molecular science field may be analyzed using constructions known as Disconnectivity Graphs (DGs) [30] and/or Barrier Trees (BTs) [32]. DGs connect “super-basins” at each subdivided energy level range. For example, evidently in Fig. 3, the energy basin can be divided into three super-basin giving rise to 3 main branches in the DG. BTs are similar to DGs, as illustrated in Fig. 3. To generate BTs, one starts in a local minimum on the energy landscape and calculates the minimal barrier height that needs to be overcome in order to reach the next neighboring local minimum. A branching point will occur at this height connecting the 2 local minima (nodes) on the tree. A rough energy landscape is represented by Fig. 3c and the global funneled nature of the more natural protein folding landscape is represented by Fig. 3b. [30].

We carried out BT analysis for the Pareto optimal sets obtained by the SPEA2-hybrid and SMS-EMOA. An aggregated energy landscape needs to be defined over the Pareto frontier’s vector space, and we used the cost function, C , defined with $w_1 = w_2 = 0.5$ in Eq. 2. Scores of sequence intermediates among adjacent non-dominated solutions were calculated assuming a series of single triplet swaps – a total of 120 steps corresponding to the number of triplets in the 4-level case. The resultant calculations are depicted in Fig. 5, where it is evident that the barriers among adjacent solutions are very high and would require sampling of a significant number of poor-scoring states to identify new solutions. The resulting “rough” energy landscape is highly sensitive to sequence variations, which renders solvers’ steps toward improvements very difficult.

BTs provide a facile way of visualizing underlying barrier heights between a adjacent local minima, which in turn allows to estimate the temperature parameters for SA and RE. However, BTs do not provide the exact landscape information about the energy basin size (e.g., the width of the valley). The presence of Pareto points hidden among “small basins of attraction”, with high barriers enclosing the basin, would render SA’s attempts to locate the global minimum extremely hard [20]; such attempts may still be successful in a Palm Tree structure, but not in a Banayan Tree structure (Fig. 3). In the case of RE, the high temperature counterpart should be set to overcome large barriers, while the low temperature counterpart should allow for effective sampling of variations within the “valley” regions [18].

Understanding how the effective performance of EAs, such as GAs and MIES, is affected by the landscape structure constitutes a harder challenge due to the complex nature of their operators. In the traditional GA, the crossover operation facilitates exploration,

also by overcoming high energy barriers, versus the mutation operator which facilitates exploitation by fine-grained search. In practice, investigating this dynamic competition with respect to the underlying (aggregated) energy landscape is a hard problem, and yet, several measures to quantify the problem hardness for EAs have been put forward [19]. Among the recently defined measures, we consider the Negative Slope Coefficient (NSC) approach [29], which estimates the empirical likelihood of objective function improvements when starting from a given search-point. An estimated NSC is computed herein for the attained Pareto optimal set and presented in Fig. 6, where fitness is defined as the negative value of the cost function value. First, a fitness cloud is visualized by plotting the Pareto frontier’s fitness values versus the average fitness of their neighbors. The average fitness value of 1000 neighbors that are n triplets away ($n = 1 \dots 60$), for each Pareto point, is calculated and plotted on the ordinate against the corresponding Pareto point’s fitness value plotted on the abscissa. Finally, this generated fitness cloud is divided into bins – such that each bin holds a single Pareto point. The adjacent averages’ cost values were then connected by short line segments. The sum of all line segments that had a negative valued slope would give us the NSC. From Fig. 6, it is evident that the current aggregated search landscape under study poses a hard problem for any search algorithm utilizing standard evolutionary operators.

5 DISCUSSIONS

Among the 3 singleobjective heuristics, RE performed better than SA and the GA; we explained that by the high temperature replica mechanism that allows extensive sampling of poorly scoring intermediates. SMS-EMOA using a MIES solver was able to obtain a better Pareto frontier in this rough landscape. Due to the dependence of the algorithm on evolving dominated intermediates toward the non-dominated solution points, multiple independent runs were necessary to obtain high coverage of the Pareto frontier.

Although both evolutionary and non-evolutionary global optimization approaches have been studied in seclusion, a direct comparison of them has seldom been performed before, especially in the current protein design multiobjective context. We have observed a *speed versus accuracy trade-off* – with RE possessing good conformational exploration and faster convergence, versus SMS-EMOA, providing better Pareto accuracy by fine-grained search. Granting additional objective function calls consistently resulted in a similar trade-off. A common high-level protocol for landscape analysis has been proposed. Particularly, for (i) estimating the energy barrier height and hence the roughness of the underlying landscape for non-evolutionary approaches, and (ii) estimating the evolvability of pathways among neighboring data points for evolutionary approaches. Landscape analysis for the bi-objective case of stability versus specificity competition – which is one of the most common formulations for protein and peptide design, was also reported. The NSC measure over the Pareto optimal set was established. The observed sequence diversity was low for the Pareto optimal set obtained by the SMS-EMOA, which could be remedied by conducting multiple runs with various setting to undergo an overall Pareto sorting, or by incorporating genetic diversity as an independent objective [28].

Figure 5: Barrier Landscape. a) Two-dimensional landscape for the 4th algorithm, indicating the minimum barrier height that needs to be overcome to move from one Pareto point to the next. b) Two-dimensional landscape for the SMS-EMOA algorithm. c) Barrier tree for 4th algorithm d) Barrier Tree for SMS-EMOA. A single search path traced from one Pareto optimal point to the next through a consecutive series of 120 single triplet mutation swaps is carried out between the search path's starting and ending Pareto points. The trace of 1000 such "triplet mutational" paths is used to find the minimal cost function barrier height that needs to be overcome to move from one minimum (point residing on the Pareto frontier) to the next. The barrier tree is constructed next. The resultant BT indicates a rough energy landscape. It also indicates that the landscape has both high and low barriers making the setting of the temperature parameter difficult for the non-evolutionary approaches (SA, RE).

5.1 Summary

Evolution strategies have been shown to perform as good as other state-of-the-art techniques on this combinatorial optimization problem (see, e.g., [24]). The 4-level heterotrimer design is a valuable test-case for algorithms development for multiobjective protein design. Despite the success of previous *forward design* approaches, this test-case supports the assertion that stability and specificity objectives are in competition, and that this conflict is exacerbated in systems of higher complexity. Where 1-level solutions were rapidly achieved using singleobjective non-evolutionary approaches [34], addressing the 4-level case resulted in high energy barriers among adjacent non-dominated solutions and low evolvability, frustrating both evolutionary and non-evolutionary approaches. Despite these challenges, the use of RE for sequence design may have promise as it allows efficient exploration of poorly scoring intermediates to traverse between optimizers constituting the Pareto optimal set. Not surprisingly, EMOAs that explicitly target non-dominated solutions perform even better. With increasing complexity, the energy landscape becomes significantly rougher, rendering optimization very hard.

Identifying the Pareto optimal set is a potentially powerful new perspective in protein design. For cases where the relative significance of multiple objectives is not clear, a set of non-dominated solution points that span the extreme boundaries of the Pareto frontier are valuable starting points for experimental characterizations.

The results of such experiments could be used to constrain the cost function or the number of objectives for subsequent generation of useful molecules. Real-world applications would require multiobjective optimizations over multiple simultaneous targets. It will be critical to continue developing algorithms that make such challenging optimization problems tractable. An idea for future work is the integration of RE into SMS-EMOA.

REFERENCES

- [1] Thomas Bäck. 1996. *Evolutionary Algorithms in Theory and Practice*. Oxford University Press, New York, NY, USA.
- [2] J. Bella, M. Eaton, B. Brodsky, and H.M. Berman. 1994. Crystal and molecular structure of a collagen-like peptide at 1.9 Å resolution. *Science* 266, 5182 (1994), 75–81. DOI: <http://dx.doi.org/10.1126/science.7695699>
- [3] Nicola Beume, Boris Naujoks, and Michael Emmerich. 2007. SMS-EMOA: Multiobjective Selection Based on Dominated Hypervolume. *European Journal of Operational Research* 181 (2007), 1653–1669.
- [4] F. Calvo. 2009. Non-genetic global optimization methods in molecular science: An overview. *Computational Materials Science* 45, 1 (2009), 8 – 15. DOI: <http://dx.doi.org/10.1016/j.commatsci.2008.03.059>
- [5] David E. Clark (Ed.). 2000. *Evolutionary Algorithms in Molecular Design*. Wiley-VCH, Weinheim.
- [6] Indraneel Das and John Dennis. 1997. A closer look at drawbacks of minimizing weighted sums of objectives for Pareto set generation in multicriteria optimization problems. *Structural Optimization* 14 (1997), 63–69.
- [7] Kalyanmoy Deb. 2015. *Springer Handbook of Computational Intelligence*. Springer Berlin Heidelberg, Berlin, Heidelberg, Chapter Multi-Objective Evolutionary Algorithms, 995–1015. DOI: http://dx.doi.org/10.1007/978-3-662-43505-2_44
- [8] William F. DeGrado, Christopher M. Summa, Vincenzo Pavone, Flavia Nastro, and Angela Lombardi. 1999. De Novo Design and Structural Characterization

Figure 6: Calculating NSC: The middle 9 of the 11 indexed Pareto points (Fig. 5b, 5d) are defined here as the reference Pareto points belonging to the children generation. Fitness is defined as the negative value of the cost function value. The neighbors of these “reference Pareto points” n triplet mutations away are generated by replacing from 1...60 random triplets of the reference Pareto point with triplets from adjacent Pareto point on the left, and 1...60 random triplets with triplets from the adjacent Pareto point on the right. By construction, these sets of 120 neighbors are “ n triplets away”, $n = 1 \dots 60$, with respect to the Pareto optimal set (note that their fitness function values are already calculated for Fig. 5). This process is repeated 1000 times, and the 120 recorded averages of these 1000 fitness values are plotted in green against the fitness value of the reference Pareto point. The average fitness values of all 120×1000 neighbors for each Pareto point is then calculated (red) and plotted against its reference Pareto point. Evidently, the slope between 740 and 742 is steep, implying that the NSC will be very low (i.e., low evolvability for EAs). Each bin is sized such that it holds a single Pareto point. It’s important to note here that the barrier height existing between any pair of Pareto points in (Fig. 5) will stay the same independent of the numbered ordering of the Pareto points (For example height between points 2 and 10 in Fig. 5 will be the same whether we calculate it by moving from point 2 to 4 to 10 or 2 to 1 to 10) since the underlying sequence orthogonality will still be the same. NSC is also robust to the ordering of the indexed Pareto points or the selection of the 2 parent sequences for cross-over since the slope in Negative Slope Coefficient is dependent upon overall genetic similarity of the entire set of Pareto points and not just neighboring Pareto point derivatives. This fact is evident when, for example, one notices that the relatively low slope among points 7, 8, 9 and 10 is not significant to the overall NSC slope calculation (NSC being in-turn representative of GA evolvability difficulty), and that their relative ordering does not make a difference.

of Proteins and Metalloproteins. *Annual Review of Biochemistry* 68 (1999), 779–819.

- [9] David J. Earl and Michael W. Deem. 2005. Parallel tempering: Theory, applications, and new perspectives. *Phys. Chem. Chem. Phys.* 7 (2005), 3910–3916. Issue 23. DOI: <http://dx.doi.org/10.1039/B509983H>
- [10] Jorge A. Fallas and Jeffrey D. Hartgerink. 2012. Computational design of self-assembling register-specific collagen heterotrimers. *Nature Communications* 3

- (2012), 1087.
- [11] David Fleishman, Sarel and Baker. 2012. Role of the Biomolecular Energy Gap in Protein Design, Structure, and Evolution. *Cell* 149 (2012), 262–273. Issue 2.
- [12] Bernd Hartke. 2004. Application of Evolutionary Algorithms to Global Cluster Geometry Optimization. *Structure and Bonding* 110 (2004), 33–53.
- [13] H.W. Hellinga and F.M. Richards. 1994. Optimal sequence selection in proteins of known structure by simulated evolution. *Proc. Nat. Acad. Sci. U. S. A.* 91 (1994), 5803–5807. Issue 13.
- [14] Po-Ssu Huang, Scott E. Boyken, and David Baker. 2016. The coming of age of de novo protein design. *Nature* 537 (2016), 320–327. Issue 7620.
- [15] Brian Kuhlman, Gautam Dantas, Gregory C. Ireton, Gabriele Varani, Barry L. Stoddard, and David Baker. 2003. Design of a Novel Globular Protein Fold with Atomic-Level Accuracy. *Science* 302, 5649 (2003), 1364–1368. DOI: <http://dx.doi.org/10.1126/science.1089427>
- [16] Andrew Leaver-Fay, Ron Jacak, P. Benjamin Stranges, and Brian Kuhlman. 2011. A Generic Program for Multistate Protein Design. *PLoS ONE* 6, 7 (07 2011), 1–17. DOI: <http://dx.doi.org/10.1371/journal.pone.0020937>
- [17] Yaohang Li, Charlie E.M. Strauss, and Andrey Gorin. 2004. Parallel tempering in rosetta practice. In *Proceedings of International Conference on Bioinformatics and its Applications (ICBA '04)*.
- [18] J. Machta. 2009. Strengths and weaknesses of parallel tempering. *Phys. Rev. E* 80 (Nov 2009), 056706. Issue 5. DOI: <http://dx.doi.org/10.1103/PhysRevE.80.056706>
- [19] Katherine M. Malan and Andries P. Engelbrecht. 2013. A survey of techniques for characterising fitness landscapes and some possible ways forward. *Information Sciences* 241 (2013), 148 – 163. DOI: <http://dx.doi.org/10.1016/j.ins.2013.04.015>
- [20] Enzo Marinari and Giorgio Parisi. 1992. Simulated tempering: a new Monte Carlo scheme. *EPL (Europhysics Letters)* 19, 6 (1992), 451.
- [21] Michael Masin and Yossi Bulchin. 2008. Diversity maximization approach for multiobjective optimization. *Operations Research* 56, 2 (2008), 411–424.
- [22] Vikas Nanda and Ronald L. Koder. 2010. Designing artificial enzymes by intuition and computation. *Nature Chemistry* 2 (2010), 15–24. Issue 1.
- [23] Lucas Gregorio Nivón, Rocco Moretti, and David Baker. 2013. A Pareto-Optimal Refinement Method for Protein Design Scaffolds. *PLoS ONE* 8, 4 (2013), e59004.
- [24] OpenAI. 2017. Evolution Strategies as a Scalable Alternative to Reinforcement Learning. *Commun. ACM* 60, 4 (2017).
- [25] Douglas H. Pike and Vikas Nanda. 2015. Empirical estimation of local dielectric constants: Toward atomistic design of collagen mimetic peptides. *Peptide Science* 104, 4 (2015), 360–370. DOI: <http://dx.doi.org/10.1002/bip.22644>
- [26] Edgar Reehuis and Thomas Bäck. 2010. Mixed-integer evolution strategy using multiobjective selection applied to warehouse design optimization. In *Proceedings of the 12th annual conference on Genetic and evolutionary computation (GECCO '10)*. ACM, New York, NY, USA, 1187–1194. DOI: <http://dx.doi.org/10.1145/1830483.1830700>
- [27] Günter Rudolph. 2012. *Handbook of Natural Computing: Theory, Experiments, and Applications*. Springer-Verlag, Berlin-Heidelberg, Germany, Chapter Evolutionary Strategies, 673–698.
- [28] Andrea Toffolo and Ernesto Benini. 2003. Genetic Diversity As an Objective in Multi-objective Evolutionary Algorithms. *Evol. Comput.* 11, 2 (May 2003), 151–167. DOI: <http://dx.doi.org/10.1162/106365603766646816>
- [29] Leonardo Vanneschi, Andrea Valsecchi, and Riccardo Poli. 2009. Limitations of the Fitness-proportional Negative Slope Coefficient As a Difficulty Measure. In *Proceedings of the 11th Annual Conference on Genetic and Evolutionary Computation (GECCO '09)*. ACM, New York, NY, USA, 1877–1878. DOI: <http://dx.doi.org/10.1145/1569901.1570212>
- [30] David J. Wales and Mark A. Miller. 1998. Archetypal energy landscapes. *Nature* 394 (1998), 758 – 760. Issue 6695.
- [31] Shira Warszawski, Ravit Netzer, Dan S. Tawfik, and Sarel Fleishman. 2014. A “Fuzzy”-Logic Language for Encoding Multiple Physical Traits in Biomolecules. *Journal of Molecular Biology* 426, 24 (2014), 4125 – 4138. DOI: <http://dx.doi.org/10.1016/j.jmb.2014.10.002>
- [32] Michael T. Wolfinger, Sebastian Will, Ivo L. Hofacker, Rolf Backofen, and Peter F. Stadler. 2006. Exploring the lower part of discrete polymer model energy landscapes. *EPL (Europhysics Letters)* 74, 4 (2006), 726.
- [33] Fei Xu, Teresita Silva, Mihir Joshi, Sohail Zahid, and Vikas Nanda. 2013. Circular Permutation Directs Orthogonal Assembly in Complex Collagen Peptide Mixtures. *The Journal of Biological Chemistry* 288 (2013), 31616–31623. Issue 44.
- [34] Fei Xu, Sohail Zahid, Teresita Silva, and Vikas Nanda. 2011. Computational Design of a Collagen A:B:C-type Heterotrimer. *Journal of the American Chemical Society* 133 (2011), 15260–15263. Issue 39.
- [35] E. Zitzler, M. Laumanns, and L. Thiele. 2001. SPEA2: Improving the strength pareto evolutionary algorithm for multiobjective optimization. In *Evolutionary Methods for Design Optimization and Control with Applications to Industrial Problems*, K. C. Giannakoglou, D. T. Tsahalis, J. Périaux, K. D. Papailiou, and T. Fogarty (Eds.). International Center for Numerical Methods in Engineering, 95–100.