Current Topics in
Peptide & Protein
Research
Vol. 18, 2017

Original Communication

From finance to molecular modeling algorithms: The risk and return heuristic

Immanuel Lerner[1], Amiram Goldblum[1,2], Anwar Rayan[3,4], Alexandra Vardi[1] and Amit Michaeli[1,*]

[1]Pepticom Ltd.; [2]Institute for Drug Research, The Hebrew University of Jerusalem, Jerusalem;
[3]Drug Discovery Informatics Lab, Qasemi-Research Center, Al-Qasemi Academic College,
Baka El-Garbiah 30100; [4]Institute of Applied Research, Galilee Society, Shefa-Amr 20200, Israel.

ABSTRACT

While machine learning techniques have greatly increased molecular modeling capabilities, the frequent reliance on stochastic algorithms is a limiting factor due to slow optimization processes. Faster algorithms are thus sought for large scope projects. Nevertheless, stochastic algorithms also provide a distinct advantage by providing a solution ensemble rather than a single optimal solution. Producing a large ensemble of solutions is critical in problems where not all solution aspects are predictable and unpredictable properties may be key to ultimate success. Similar problems have been tackled previously before the advent of machine learning, in the field of finance. In 1952, Harry Markowitz introduced the modern portfolio theory (MPT), which uses the heuristics of risk and return to optimize a financial portfolio. In this study we will introduce an implementation of MPT heuristics in the field of protein-protein and peptide-protein interface design and show examples of its usage.

KEYWORDS: modern portfolio theory (MPT), risk adjusted design (RAD), stochastic dominance (SD), computer design, heuristic.

INTRODUCTION

The design of novel protein-protein and peptide-protein interfaces holds vast potential in medicinal and agricultural biologics and remains a key goal for computational molecular design. Two key challenges are faced when attempting to accomplish this goal: An enormously large search space, requiring unrealistic computation times and flawed computational prediction methods that may “miss” many good solutions [1]. Overcoming these problems is critical for computational peptide/ protein design.

Excessively large search spaces inhibit reaching optimal solutions by exhaustive search, as searches will not be performed within a realistic time frame. To overcome this, numerous stochastic algorithms have been developed. These methods began with rather simple Monte Carlo simulations that direct the random search space from the current step [2]. As time progressed, more advanced learning methods such as iterative stochastic elimination (ISE) [3] and genetic algorithms (GA) [4] were introduced. These methods generally follow the process of random sampling, followed by elimination of unfavorable parameters and/or the amplification of favorable ones. This allows for a much more time realistic solution convergence, although the need for stochastic sampling still remains computationally cumbersome. Algorithms that can also be applied deterministically, such as dead-end elimination (DEE) [5] have the theoretical ability to speed this process. However, due to the strict elimination criteria, they are often used in conjugation to stochastic algorithms [6]. As of today, mainstay design algorithms all

*Corresponding author: amit.michaeli@pepticom.com

use stochastic sampling, which is arguably the rate limiting step towards convergence.

The ability to calculate binding affinities *in silico* has not been perfected, serving as another obstacle to protein/peptide design. For example, the renowned ROSETTA energy function, used in numerous successes in protein and peptide design [7-9], was analyzed on several complexes and shown to have an overweighed emphasis on electrostatic over hydrophobic interactions [10]. An interesting approach towards tackling the shortcomings of these scoring functions focused on a multivariate energy funnel analysis to discern near-native from other conformations during docking [11], improving scoring resolutions indirectly. Nevertheless, a layer of uncertainty on computational protein/peptide design scoring still remains, with a significant portion of binding biophysics remaining unaccounted for. Scoring imperfections effectively eliminate the goal of reaching the single "global minimum" solution, as proposed solutions are merely hypotheses, with increased probabilities of outperforming randomly generated models. Solution ensembles, provided by most stochastic algorithms hence have an advantage, offering a set of hypotheses as opposed to a single one.

Interestingly, analogous problems were encountered in the field of finance during the middle of the last century while exploring methods for investment portfolio optimization. Investment portfolios are composed of numerous liquid assets, of which each can be held at variable portions, creating an enormous search space, analogous to the combinatory search space in peptide/protein design. Similarly, the ability to predict the future performance of these individual assets is rather poor, creating the need for investment in a portfolio/ensemble of assets rather than a single one. The modern portfolio theory (MPT) or mean-variance analysis [12] was introduced in 1952, by Prof. Harry Markowitz who was awarded the Nobel Prize for its development in 1991. One of the fundamental concepts introduced by this theory is investor risk aversion. This implies that investors expect to be compensated for the investment risks undertaken. Investors will avoid an investment opportunity if alternatives of lower risk and similar expected return or similar risk and higher expected returns exist.

In the original paper, Markowitz used the arithmetic average of historic returns as the measurement for expected returns and the historic arithmetic standard deviation as an asset's risk measurement [12]. These were later combined with pairwise historic correlation coefficients to deterministically calculate the asset weights that produce an optimal portfolio, with the highest return per risk ratio [12]. Today, a variety of risk measurements are used in numerous optimization techniques for the accomplishment of an optimal risk-averse investment portfolio. Notably, the MPT was performed prior to the availability of high-end computational technology and was performed without stochastic sampling. MPT-based optimization approaches have since been adopted to irrigation water management [13], biodiversity conservation [14], climate change management [15] and even national security [16].

In this paper we will introduce an MPT approach for rigid-body side-chain design in protein-protein and peptide-protein interfaces and demonstrate a few theoretical and lab-validated design cases. In these cases, the input structure contains a pre-positioned interface of poly-glycine, which was designed to bind the partner protein using all rotamers for all 20 naturally occurring amino-acids. We will compare the results to those obtained by Monte Carlo simulation [2], the high-end ISE algorithm [3] and the fast but complex problem incompatible random-greedy algorithm.

MATERIALS AND METHODS

Protein structure treatment

Protein-protein and protein-peptide complexes of resolution 2.5Å or better were selected. The complexes were then checked using the MolProbity server [17] (http://molprobity.biochem.duke.edu/). No protein structures with missing any interfacial or backbone atoms were taken. Protein structures with missing side-chain atoms were corrected by sampling the rotamers [18] of the appropriate residue and selecting the lowest potential energy rotamer, calculated using the AMBER force-field [19], with the GB/SA solvation model [20, 21]. Hydrogen atoms were explicitly added with the most common ionization state at pH7 (charged Lysine, Arginine, Aspartate, Glutamate and Histidine along with N and C termini).

In interfacial side-chain optimizations, the interface residues to be optimized (designed) were defined as residues with an inter-protein backbone Cα-Cα atom distance of less than 16Å, to at least one of the partner protein's Cα atoms. Only residues that fit this criterion were considered for optimization.

When performing mutations to optimize the binding to a partner protein a risk of destabilizing the designed protein exists. To prevent the destabilization of the folded structure of the protein/peptide being optimized, the protein's/ peptide's conformational energy potentials were calculated. Interface residues that contributed -1.5 Kcal/mol, or lower were excluded from mutation; this is a user defined threshold (selected by the individual user) and can be repeated with different cutoffs.

Expected "Returns" in side-chain optimization

To use the risk-return heuristic we must first define risk and return in protein design. Most molecular mechanics force fields and other energy scoring functions are composed of the summation of pairwise terms. Under this assumption we can test the binding affinity of each rotamer in each possible position, with all other designed positions mutated to Glycine. This energy (minus the backbone interaction energy) represents the rotamer's marginal contribution to a hypothetical complex. This marginal contribution is henceforth labeled **E_X(i)**, or the expected return of a particular rotamer i in position X (a position is normally the residue number in the modeled chain). It is important to note that this is an approximation, as physical interactions that are not pairwise are known to exist. For example, a residue involved in hydrogen bonding would show a higher marginal energy contribution if surrounded by bulky residues that increase its effective Born radius (lowering the dielectric constant of the electrostatic component dominant in a hydrogen bond), in a concurrence to the "O-ring" structures observed by Bogan and Thorn [22]. It is for that reason that we label the return for choosing rotamer i in position X as expected, as modifications in the final model may lead to some deviations from this value. This, and other force-field inaccuracies are similar in nature to the usage of a stock's historic average return as its expected return in the future, a common practice in financial optimizations using the modern portfolio theory and which is of course, also probabilistic in nature.

"Risk" in side-chain optimization

The "risk" in choosing a particular rotamer is attributed to rotamer options being mutually exclusive with others. In its most robust form, this is caused by two rotamers in different positions that occupy the same Cartesian space, leading to atomic clashes. Other forms of mutual exclusivity such as electrostatic repulsion, lack of backbone conformation compatibility and solubility also exist, but are not included in this study for simplicity's sake. In cases where the highest return rotamers in all positions, are risk-free, the modeling problem is reduced to sampling, scoring and picking the highest return rotamer in each position- a "Greedy Algorithm". Mutual exclusivities, or "risk" prevent the fast "Greedy Algorithm" deployment, requiring an algorithm compatible with higher complexities.

For example, assume a simple system of two positions X and Y being optimized to bind a target protein. Each position has a "hot-spot" amino-acid rotamer; i for X and j for Y and both positions have a "risk-free", '(rf) alanine residue with E_Y(rf) = E_X(rf) = 0 Kcal/mol. Rotamer i has the highest expected return in position X so that E_X(i) = -5 Kcal/mol, but also clashes with the superior return rotamer, j of position Y. where E_Y(j) = -7 Kcal/mol. A "Greedy Algorithm" that would sample X first, would be barred from choosing rotamer j in position Y and would have to choose the rf rotamer, leading to a final binding energy of -5 Kcal/mol. In this case an optimal solution would be the selection of j in position Y and rf in position X, to create a final binding energy of -7 Kcal/mol. At this stage rotamer i of position X has both the qualities of "risk" and "return". Selecting it would contribute -5 Kcal/mol towards binding, but at the "risk" of loosing a contribution of -7 Kcal/mol of j in position Y. Likewise, rotamer j of position Y has a return of -7 Kcal/mol and "risk" of -5 Kcal/mol.

Another important aspect of measuring the "risk" in picking a rotamer is the cost of substitution of the rotamers with which it is mutually exclusive, in the same position. Taking the above example where $E_X(i)$ = -5 Kcal/mol is mutually exclusive with $E_Y(j\ )$ = -7 Kcal/mol, if we introduce a new rotamer, k as the best non-mutually exclusive energy contributing rotamer in position Y, so that $E_Y(k)$ = -6, then the cost of selecting i on position X, due to position Y is (-6)-(-7) = 1 and depending on its interactions with other positions, i could still be favorable for the optimum solution. However if $E_Y(k)$ = -1, the cost of selecting i on position X due to Y is (-1)-(-7) = 6. It is hence intuitive that the mutual exclusivity with a "hot-spot" rotamer in the case where numerous "hot-spot" rotamers are available in the same position is less "risky" than mutual exclusivity with the same rotamer energy without other potential "hot-spots" in that position. This substitution factor can be shown as the absolute value of the energy contribution of a rotamer divided by the sum of the absolute values of each contributing rotamer in the same position, denoted as $P_Y(j)$ for rotamer j in position Y:

$$py(j) = \left(\left|\Delta Ey\,(j)\right|\right) / \sum_{t=0}^{N} \left(\left|\Delta Ey\,(i)\right|\right)$$

It is therefore clear that the risk measurement of selecting a particular rotamer is composed of both the returns of mutually exclusive rotamers, as well as their marginal cost of substitution. We devised the following definition of risk of selecting rotamer i in position X:

$$Rx(i) = \sum_{x \neq Y}^{N} \sum_{j=l}^{j=n} (Bij)\,(Py(j))\,(\varDelta Ey(j))$$

where Bij is a Boolean operator that takes the value of 1 when i and j are mutually exclusive and 0 for all others, $P_Y(j)$ is the relative energy contribution of the rotamer j in position y and ΔEY(j) is the "return" of j. Rotamers that show returns poorer than glycine are not included in the risk measurement, as for any given position, glycine serves as a risk-free asset. Since glycine has no side chain, it cannot be mutually exclusive with side chains from other positions. Thus, rotamers that provide expected returns more positive than or equal to glycine are by definition inefficient, offering no compensation for the risk taken by choosing them. There can also be riskless rotamers other than glycine that shows no mutual exclusivity. In such cases, glycine may be rendered "inefficient" and eliminated.

Risk adjusted design (RAD) optimization procedure

Prior to optimization, the energy return for each possible rotamer in each possible position is calculated. Rotamers that add a more positive interaction energy than glycine or clash with the designed sequence's backbone are automatically eliminated at this stage. During this procedure, every tested rotamer "signs" a grid (automatically determined to include all space within the cutoff distance) at every grid point which is inside its atomic radii. Rotamers which "sign" the same grid point are considered mutually exclusive and have a risk Boolean (Bij) of 1 to one another.

The optimization procedure is performed iteratively. At the beginning of each procedure the risk of each rotamer is evaluated as described above. For each position, each possible rotamer is checked against all other rotamers in that position. Rotamers which have a higher risk without the compensation of excess returns from (or alternately, lower returns at the same risk) any other rotamer in that position are added to the inefficiency score. In each iteration, the rotamer with the highest inefficiency score is eliminated. It is important to note that the elimination of a rotamer changes the risk estimate of all rotamers with which it has a Bij of 1, The elimination also changes the probability substitution weight $P_Y(j)$ in the position of the eliminated rotamer. At the end of the elimination process, an "efficient set" of rotamers remains in every position and the "risk" values are re-calculated. The iterations continue until no more elimination can take place, because all rotamers are "efficient", or alternately, the combinatorial size allows for an exhaustive search of the remaining rotamers.

Iterative stochastic elimination (ISE)

Iterative stochastic elimination (ISE) has been previously described [3], with successful applications in protein flexible fragment conformational searches [3, 23], cyclic peptide design [24], ligand docking [25], and chemoinformatic models [26]. Briefly, ISE is a generic discrete combinatorial optimization method that, in its application to finding the global energy minima, iteratively eliminates values which contribute consistently to highest energy conformations, and to a lesser extent to lowest energy conformations. The algorithm begins by constructing a matrix that contains a set of the possible (discrete) values for each degree of freedom (variable) that defines the problem (system). If the problem is molecular conformation/composition and the degrees of freedom are amino acid rotamers, the dihedral angles for each rotamer represent individual variable values. One rotamer is randomly picked for each interface position and is compatible with previously picked rotamers in other positions. After all positions are picked, the protein/peptide binding affinity is calculated as described above. This step is repeated many times to form a large sample, usually in the range of 30,000 sampled full conformations. The scores of that sample are arranged in a virtual histogram in which only a small fraction (1-10%) of worst and of best results are examined in detail and compared to assess the contribution of each and every variable value to the final scores of these best and worst results.

A value that appears in the worst results with a significantly higher frequency than expected from its random distribution (based on its total appearance in the full sample) and appears with a significantly lower frequency than expected among the best results, is marked for elimination. The next iteration of random picking, scoring, sampling, and eliminating thus begins with a smaller number of possible combinations. The elimination process is performed iteratively until the number of possible conformations enables exhaustive search in feasible time.

Monte Carlo side-chain optimization (MC)

Monte Carlo sampling was performed as follows: The designed peptide/protein’s interface was mutated to Glycine. As a random selection cycle is initiated, the interaction energy at the current state, s0, is calculated. A random rotamer is then selected at a random position and the energy at the new state, s1, is calculated. An acceptance probability p(s1) of the new state is then calculated as follows:

$$p(s1) = e^{((s0\text{-}s1)/RT)}$$

where, e is Euler’s number, R is Boltzmann’s constant and T is temperature in degrees Kelvin. This probability is compared to a random number from 0 to 1; if the probability is greater than the random number, the new state is accepted and becomes the current state for the next iteration, while a random number higher than the probability leads to a rejection of the new state and the preservation of the current state. The temperature was set as constant, at 300 °K.

Random greedy side-chain optimization (RG)

Random greedy optimization was performed as follows: The designed peptide/protein’s interface was mutated to Glycine. Positions were then chosen in random order and the lowest energy rotamer was selected for each position. At each position, only rotamers compatible with the backbone and rotamers previously selected are screened. The optimization is run in parallel, with each process following a random position section order.

RESULTS

We selected several test cases to demonstrate the risk adjusted design (RAD) algorithm, which implements the “risk” and “returns” criteria used in the modern portfolio theory for rigid-body interface design. For this purpose, we are describing below several theoretical and lab-validated cases.

We have previously reported about our *ab initio* discovery of a structure-stabilizing chaperone for Y329S mutated human Glycogen Branching Enzyme 1(hGBE1), leading to the ultra-orphan adult polyglucosan body disease (APBD) [27]. The Y329S mutation causes a complete or near complete loss of function of hGBE1, with patient cells showing 0-5% of normal human activity [27]. The crystal solution of the wild-type (WT) hGBE1 showed that the Y329S mutation area was solvent accessible, allowing for a potential peptide chaperone to compensate for the loss of non-polar interactions of the benzene ring and hydrogen bond of the hydroxyl group [27]. This required screening thousands of possible backbone conformations followed by side-chain optimization, with different length and size. The RAD algorithm was used for side-chain optimizations, and out of the solution ensemble one peptide, LTKE, was selected for testing. LTKE showed a K_d of 1.6 µM, and was able to partially rescue about 27% of normal hGBE1 activity in patient cells [27].

The RAD optimization process of the LTKE chaperone of Y329S hGBE1 has not been described thus far and will be detailed here as an example application of RAD. Since it is a short peptide, it allows us to follow the RAD process in detail. RAD began by analyzing the initial "risk" and "return" values for all four amino-acid positions (Figure 1A). Observation of the peptide's "risk-return" graphs shows that two out of four amino acid positions, in positions 1 and 3, are "risk-free" for the best solution (Figure 1A). RAD can at this point select the leucine residue for position 1 (Figure 1A, top left) and the lysine residue for position 3 (Figure 1A, bottom left). Being "risk free" indicates that this will not adversely affect the rest of the design process and hence, there is no need to further sample these positions. When optimizing a 4-position peptide

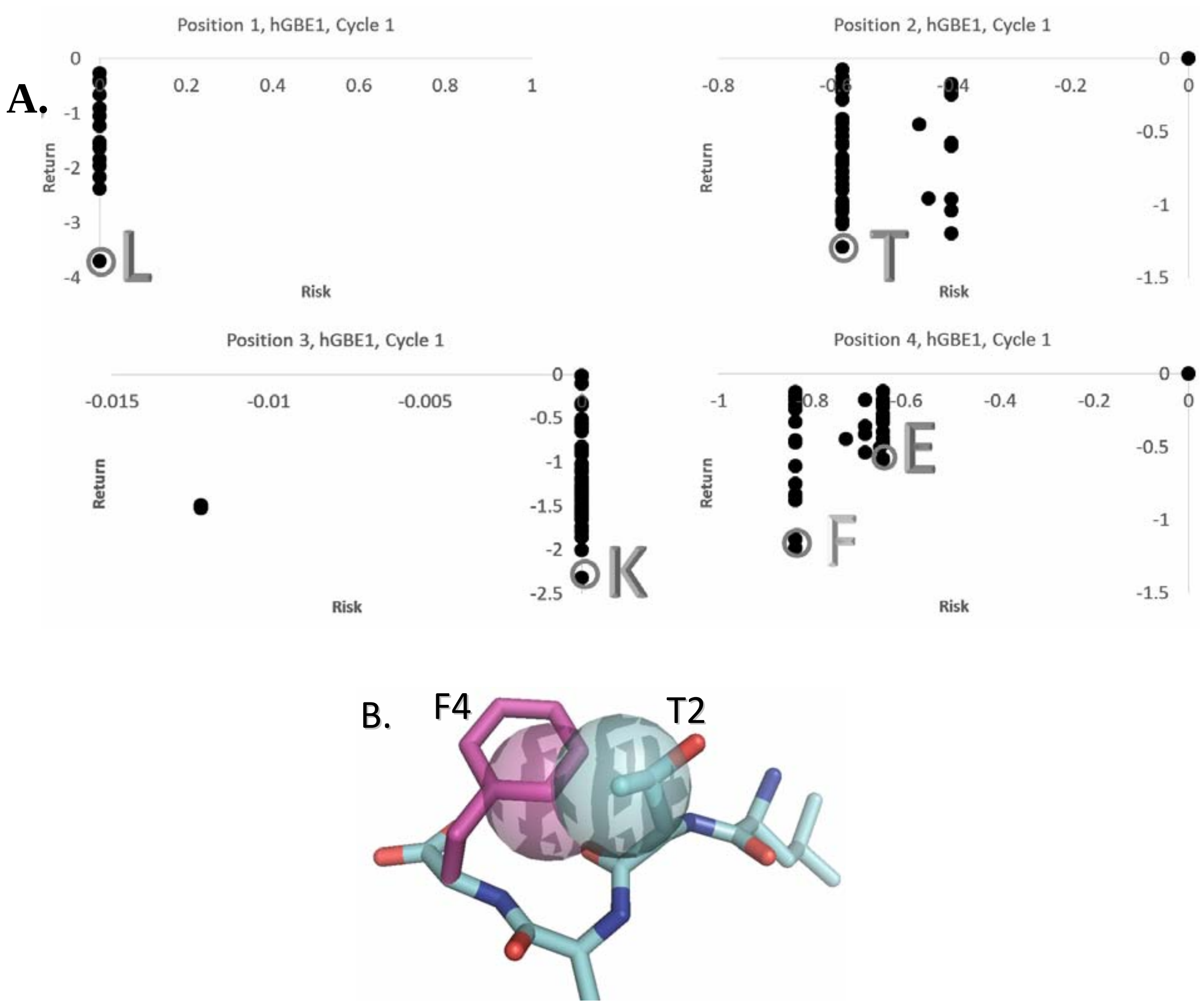

Figure 1. Initial risk and return graphs for binding to hGBE1. (A) The selected leucine rotamers for positions 1 and 3 are both of best "returns" and "risk-free", and selecting them will not adversely affect other positions. Positions two and four are "risky" requiring iterative optimization. **(B)** One risk factor was caused by an atomic clash between the best "return" phenylalanine of position 4 (purple) with threonine of position 2 (cyan), shown by the intersecting spheres.

we often presume a 20^4 = 160,000 amino acid sequence combinatory solution space. The identification and separate optimization of "risk free" best solution positions reduced the solution space to only 20^2+2(20) = 440 iterations. While for such a small peptide, both solution spaces are feasible, the latter is much faster, especially when there is a need to sample thousands of backbones, as was done in this case.

Two "risky" positions remained: Position 2, for which a "risky" threonine rotamer is also of the best return, and position 4, for which a phenylalanine rotamer was the best "return". The "risk" for both best "return" rotamers was partially due to their mutual exclusivity, caused by an atomic clash (Figure 2B). Exhaustive search determined the LTKE combination to be the optimal one in the solution ensemble, with glutamate in position 4 being compatible with the threonine in position 2. The reduced complexity was also evident when the "LTKE" peptide was attained by running a "Random Greedy" algorithm on the same backbone. In cases where position 2 was sampled randomly first, threonine was selected, which prevented the selection of phenylalanine in position 4 (Figure 2B), with the best compatible option being the glutamate rotamer.

An *ab initio* design effort was also accomplished using the RAD algorithm on peptides designed to bind to Toll-like receptor 4 (TLR4) (to be published separately; in preparation). TLR4 serves as the innate immunity receptor for lipopolysaccharide (LPS) found in gram negative bacteria. TLR4 binds LPS *via* two co-receptors: MD2 and CD14 [28]. The design effort was focused on creating a set of peptides targeting the LPS binding pockets of both the MD2 (PDB: 2Z65) [29] and CD14 (PDB: 1WWL) [30] co-receptors. Design was performed by running the RAD algorithm on multiple, randomly docked, poly-glycine amino-acid chains and searching the solution ensembles for identical "hot-spot" peptides that can bind both co-receptors. Eight identical "hot-spot" peptides were discovered for two 9 amino-acid backbone chains in a linear conformation for MD2 and helical conformation for CD14. The risk and return criteria for position one, optimized for binding MD2 (Figure 2) and CD14 (Figure 2) are analyzed here as an example. For both MD2 (Figure 2A) and CD14 (Figure 2A)

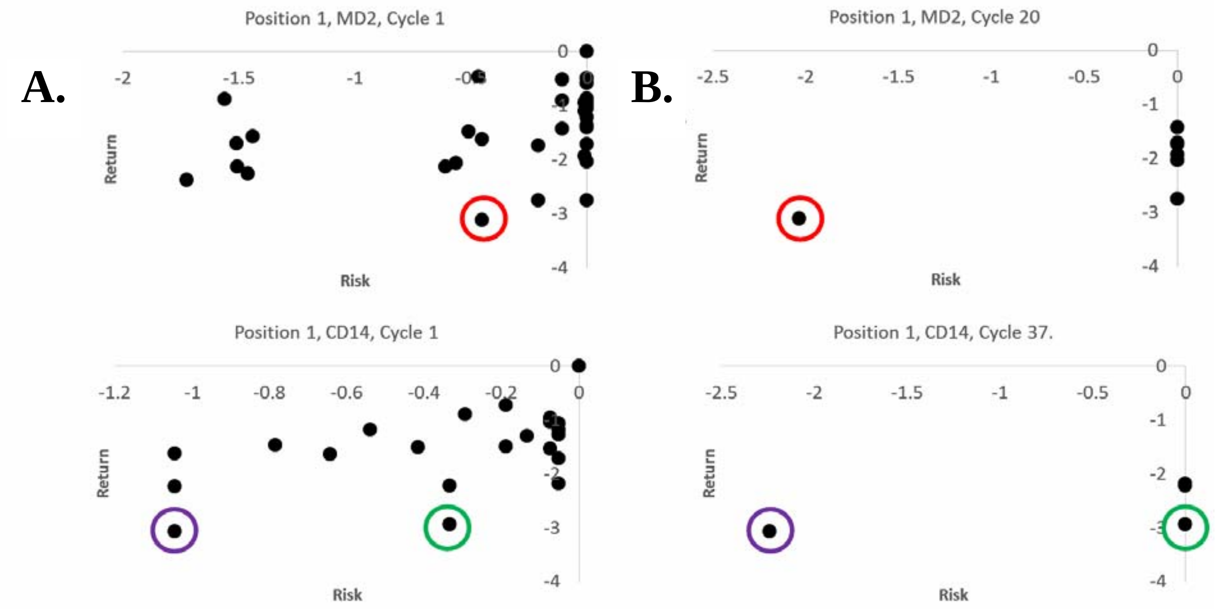

Figure 2. RAD was used to design a 9 amino-acid backbone for binding to both MD2 and CD 14 co-receptors of TLR4. (A) The risk/return graph for position 1 at cycle 1 shows different pattern for MD2 (top) and CD14 (bottom). **(B)** At the end of the elimination process a smaller number of rotamers remain. There is a change in risk from that calculated during cycle 1 for selected rotamers (**by respective colored circles in A and B)**.

position 1 was not “risk-free”. The RAD algorithm iteratively reduced the number of rotamers in the solution space from 34 (Figure 2A) to 8 (Figure 2B) for MD2 and 22 (Figure 2A) to 4 (Figure 2B) for CD14. As eliminations of residues excluded from the efficient set take place in each iteration, risk values may increase or decrease due to the changing mutual exclusivities (Figure 2A, B). In the case of position 1, the elimination cycles ended prior to “efficient set” convergence, as the selected exhaustive solution space of 5,000 was reached. Of the 9 amino-acid backbone positions, only position 9 for MD2 and positions 5 and 9 for CD14 showed a “risk free” best “return” solution. The solution space was reduced from 10^{13} (CD14) and 10^{12} (MD2) to 5,000 for each, at which point an exhaustive search was performed, producing the solution ensemble. The 8 sequence identical ensemble was biologically validated, with 2 peptides showing activity on both MD2 and CD14, one only on MD2 and 5 showing no activity on either [31] (unpublished results).

To demonstrate more potentially combinatorial problems, we performed theoretical interface re-design for protein-protein interfaces, for which one of the protein interface residues were replaced with glycine. The re-design was performed independently with RAD, ISE, MC and “Random Greedy” algorithms. The complex between Colicin Dnase E9 and its cognate immunity protein, IM9 has been the focus of numerous studies characterizing protein-protein interactions [32-35]. Colicin Dnases are plasmid encoded proteins that are cytotoxic to *Eschericia-coli*, cleaving its DNA. The bacterial cell that expresses a Colicin Dnase also expresses its cognate immunity protein, protecting its own DNA *via* the formation of a tight protein-protein complex. Immunity proteins can also form complexes with non-cognate Dnases, and these complexes tend to have remarkably lower binding affinities when compared to the cognate complexes. The E9 Dnase has X-ray structures solved for both its cognate complex with IM9 [36] (PDB: 1EMV, PDB: 1BXI) and its non-cognate complex with IM2 [34] (PDB: 2WPT), with a highly conserved backbone structure (RMSD = 0.68Å). Coupled with extensive experimental binding affinity data, the E9/IM9 and E9/IM2 complexes serve as excellent models for rigid body interface design.

The Gibb’s free binding energy of the E9 Dnase with its cognate immunity protein IM9 complex has been experimentally determined at -22 Kcal/mol [37], while using our software (with AMBER GB/SA included for energy evaluations) a binding energy of -20.9 Kcal/mol was calculated. Alanine scanning mutations for this complex have been reported experimentally [35] and calculated *in silico*, using the Rosetta Alanine scanning procedure (using the Robetta server) [38, 1]. An AMBER GB/SA calculated repeat of the alanine scanning experiment by Wallis *et al.* [35] showed a correlation of 0.76 to the experimental alanine scanning mutagenesis of these rotamers.

Next, we mutated interface residues of IM9 to glycine with return and risk matrices generated by sampling all possible rotamers in each position. Following matix generation, E9 binding was optimized for each position using RAD. To exhaustively explore all rotamers remaining after the initial elimination of rotamers that have positive binding energy and rotamers clashing with the backbone, 10^{42} possible interface models remain and need to be evaluated. The RAD elimination cycles were able to reduce this number to 10^{17} following the first cycle and to a manageable 10^{5} after the fourth cycle, taking only seconds to reach this stage that enables full calculation of all 10^{5} options.

We postulated that although this was a full-sized protein-protein interface, many of the significant residues were either “risk-free” or nearly “risk-free” meaning that their mutual exclusivity with rotamers of other positions that could provide a major contribution to the binding energy is small. To demonstrate this, we ran 100 independent “Random Greedy” (RG) simulations, with each simulation covering all interface positions in a random order. The best and worst RG models had a calculated binding energy of -27.3 Kcal/mol and -24.7 Kcal/mol, respectively. While the top 100 scoring models in the ensemble of solutions produced by RAD had calculated binding energies ranging from the best with -27.9 Kcal/mol to the worst with -27.1 Kcal/mol, suggesting that this was in fact the case and that the RAD algorithm converged to greedy algorithm in numerous positions.

The top RAD models conserved the rotamers of residues L33/V37/Y54/Y55, shown experimentally to have the highest contribution to the binding energy. The valine in position 34, also shown to be experimentally significant, was replaced with a methionine rotamer that interacts similarly with a hydrophobic groove on the partner protein (Figure 3A). The E30 residue, experimentally determined to have a ΔΔG of 1.4 Kcal/mol [35] (note that in alanine scanning experiments positive ΔΔG values indicated a binding affinity contribution) and calculated to be significant both by RAD and by Robetta due to its salt bridge with the partner protein's Arginine 54 residue, was also conserved. The E41 residue, which when mutated to alanine was shown to lower complex affinity (ΔΔG) by 2.08 Kcal/mol [35], was also conserved but with a different rotamer that allows for its carboxyl group to form a stable salt bridge with E9's K97 and K89. IM9's D51 residue, which contributes a significant portion of the binding energy *via* interfacial water-mediated H-bonds was not preserved in the RAD or ISE solution ensemble, possibly due to the inability to calculate its contribution using an implicit solvation model. The S50 residue was partially conserved in the RAD and ISE solution ensembles.

In addition to the conservation of the majority of the substantial energy contributors in our modeling of the native IM9 protein, new contributors of binding energy were also found, allowing for improved binding energy by the

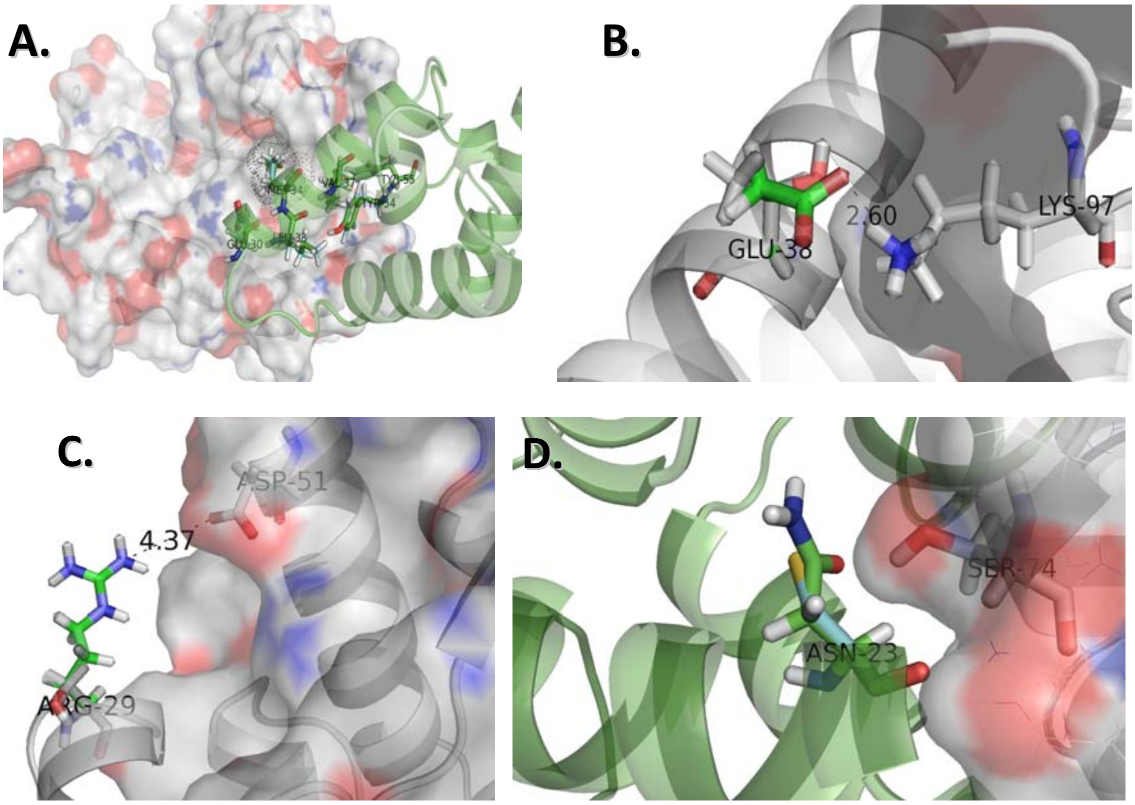

Figure 3. The redesign of IM9 interface to better bind Colicin E9 (PDB:1EMV). (A) After RAD algorithm convergence, 4 of the 5 experimentally determined (non-water mediated) most contributing residues are conserved in the designed protein (green) when compared to the WT (blue). The E30 residue is also conserved. **(B)** The WT T38 residue is replaced by E forming a salt bridge with the K97 of E9. **(C)** The S29K replacement allows for a weak electrostatic interaction with D51 on E9. **(D)** The C23N replacement forming a hydrogen bond with S74 on E9.

RAD and ISE solution ensembles. The replacement of T38 with E38 was widespread in both RAD and ISE solution ensembles. T38 was experimentally shown by alanine scanning to have a ΔΔG of 0.9 Kcal/mol [35], and calculated to have a ΔΔG of 0.012 Kcal/mol by Robetta and 0.27 Kcal/mol by us. The replacement of T38 with E38 allows it to co-interact with E41 in forming a salt bridge with E9's K97 (Figure 3B), increasing its Robetta-calculated ΔΔG to 1.9 Kcal/mol. Another example is the replacement of S29 with K29, allowing it to contribute electrostatic energy through its interaction with E9's D51 (Figure 3C). The interfacial C23 residue, experimentally found to have a ΔΔG of 0.92 Kcal/mol [35], and calculated to have a ΔΔG of -0.15 Kcal/mol by Robetta and 1.02 Kcal/mol by AMBER GB/SA was replaced by N23, calculated by Robetta to contribute a ΔΔG of 1.5 Kcal/mol, forming a hydrogen bond with E9's S74 (Figure 3D).

To further demonstrate the concept of design risk, we ran 30 separate constant temperature Monte Carlo (MC) simulations. Simulations were carried out either with one cycle of RAD prior to the MC simulation (rMC) or without. The resulting top models from each of the two sets show that the MC following one cycle of RAD led to a superior ensemble, with the worst of 30 rMC models showing better calculated affinity than the best model produced by the MC set without the RAD cycle (Figure 4).

The complex of E9 with its non-cognate IM2 shares a similar backbone to the cognate complex (backbone RMSD = 0.39, full heavy atom RMSD = 0.68), and its structure has been solved using X-ray crystallography [34] (PDB: 2WPT). The non-cognate complex has a much weaker binding affinity to E9 when compared to the cognate complex, with a Kd value that is 7 orders of magnitude higher than the cognate complex (thus, nearly 10 Kcal/mol less binding energy to that of the cognate complex) [34]. Using our energy function, we tested the marginal contributions of each of the residues. The individual residue energy contributions, serving as the "returns" in the RAD algorithm, show a correlation of 0.64 to the binding energy differences measured in experimental alanine scanning mutagenesis [34]. As previously reported using calculations with RosettaDesign [34], AMBER GB/SA failed to identify the role of D33, a residue that when mutated to alanine improves the binding affinity of the non-cognate complex by two orders of magnitude.

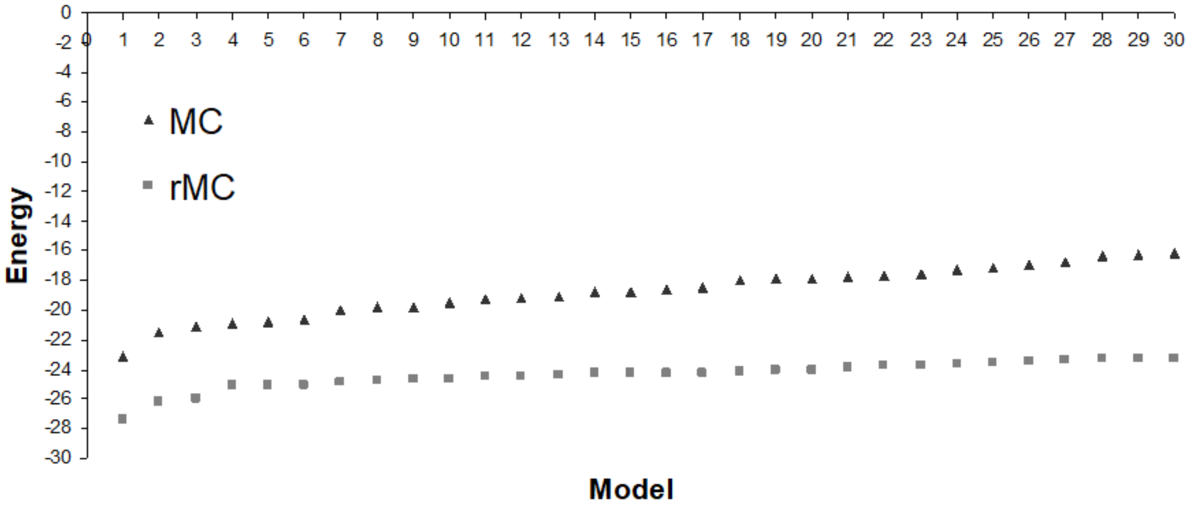

Figure 4. The effects of one cycle of RAD on a MC ensemble. 30 Monte Carlo simulations were performed on the IM9 interface (PDB: 1EMV) and displayed according to rank (X-axis) and the best calculated binding energy (Y-axis) in 40,000 steps. The simulations were independently performed, with the rotamers remaining after one cycle of RAD (rMC, grey squares) showing a better calculated affinity than the simulations conducted with all rotamers present (MC, black triangles).

The residues of the non-cognate complex partner IM2 were mutated to glycine and optimized for binding E9 using RAD. In the top 100 models of the solution ensemble, 100% of the models retained the same rotamers of V37/Y54/Y55 as in IM2, experimentally validated hot spots common to both cognate and non-cognate complexes. The D33 residue was replaced in all models, with leucine being the most common replacement, interacting with the F87 hot spot in E9 in a similar fashion to the cognate leucine in position 33 (Figure 5B). The experimental mutation of this leucine to alanine in the cognate complex lowers its binding affinity by 2 orders of magnitude while the mutation of the aspartic acid to alanine in the non-cognate increases its binding affinity by 2 orders of magnitude, validating the importance of the D33L mutation (Figure 5B). The second most common amino acid in position 33 was Q33; while no experimental evidence of this replacement has been reported, the Robetta alanine scanning server calculated that the mutation Q33A has a ΔΔG of 6.7 Kcal/mol [38, 1], mostly due to hydrogen bonding with S71 in E9 (Figure 5C). As in the cognate complex, the E30 salt bridge was retained in the RAD and ISE

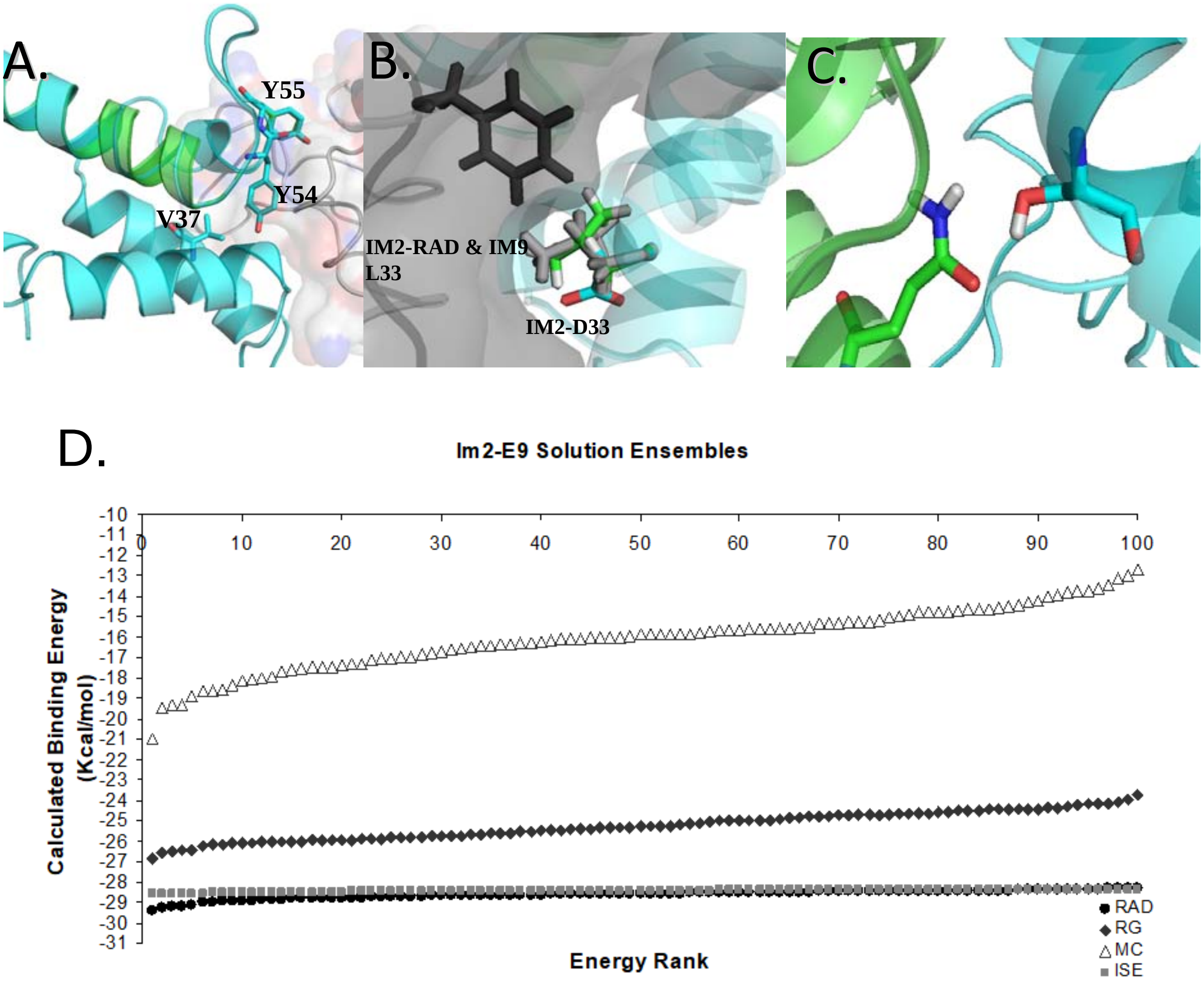

Figure 5. The redesign of IM2 interface to better bind to colicin E9 (PDB: 2WPT). (A) After completing the design process 4 of the 5 most contributing residues are conserved in the designed protein (green) when compared to the WT (blue). **(B)** The unfavorable non-cognate D33 (blue) residue is replaced by L33 (green) very close in conformation to the cognate L33 (white). **(C)** Glutamine in position 33 hydrogen bonding with serine 71. **(D)** A comparison of the top 100 models using MC, RG, ISE and RAD.

solution ensembles and E41 residue was changed to be able to form a stable salt bridge with E9's K97 and K89. The S50 residue is partially conserved and the D51 water-mediated interaction is lost and replaced with a Methionine that has a calculated contribution of 1 Kcal/mol.

As in the cognate E9 complex, optimizing the binding interface using MC yielded an ensemble of lower binding affinities. By the RAD algorithm, 100% of the solution ensemble was found to be better by 10 Kcal/mol or more than the native complex (with IM2) binding affinity. While, by the MC, only 1 of 100 experiments yielded a better complex than the native, by 1 Kcal/mol (40,000 iterations per experiment) (Figure 5D). Of the MC ensemble, only the top scoring model retained the same rotamers of V37/Y54/Y55, the experimentally validated hot spots, with the D33 residue replaced by glycine.

DISCUSSION

Risk adjusted design (RAD) implements the "risk" and "returns" criteria used in the modern portfolio theory to generate an algorithm aimed at performing side chain design optimization reaching the same quality of solutions as "high-end" algorithms such as ISE but significantly faster, due to the lack of stochastic sampling. Stochastic sampling is performed under the null hypothesis, that the interface design problem is entirely combinatorial. As such, in the case of a fully combinatorial problem, the solution space is:

$$\mathit{Solution\ Space}\ (\mathit{naïve}) = 20^{(n)}$$

representing the 20 naturally occurring amino-acids on an interface of n amino-acid positions. On these compositions, in each position, all possible rotamers need to be sampled and calculated. Under the opposite hypothesis, that the problem is entirely not combinatorial, a "Greedy" algorithm can be used, requiring 20(n) amino-acid permutations, including all rotamers, a much smaller amount than under the combinatorial null hypothesis.

The "entirely combinatorial" null hypothesis is hence a computationally expansive paradigm. Showing that when it is at least in some cases, not true, can greatly reduce search space. In the example problems shown above, the RAD algorithm was able to distinguish the interface portions for which a combinatorial hypothesis was required. In the case of a number, x of amino-acid positions where the best solution was also "risk-free" the sampling space was effectively reduced to:

$$\mathit{Solution\ Space}\ (\mathit{After}\ 1^{st}\ \mathit{RAD\ Cycle}) = 20(x) + 20^{(n-x)}$$

effectively separating the non-combinatorial component and enabling a significantly faster design, with the fast "greedy algorithm" being used to optimize the non-combinatorial portion. Additionally, the usage of matrix rather than stochastic sampling cycles for eliminations provides another speed advantage.

In the field of financial asset analysis, it has been shown that in order to better fit the investor's utility function, one investment opportunity dominates another by second-degree stochastic dominance [39]. The capital asset pricing model uses historic standard deviations as a measure of risk [12], which has been later replaced with risk models more consistent with second-degree stochastic dominance [40]. Investor utility functions show concavities which best fit second-degree stochastic dominance outcome distributions, due to investor risk aversion [41]. Given two possible outcomes, where one provides a certain return and the other provides the same return with an added random "noise", (which can be equally positive or negative) investors will choose the former over the latter [41]. This investor utility is understandable due to the nature of long term investments, where total gain is calculated by a geometric average, so that a percentage of gain does not compensate for the same percentage of loss. Intuitively, the utility function for scientists searching for new molecules is dissimilar, as the utility function is determined by the properties of the best molecule discovered and not by the distribution properties of the entire screened set. If this is indeed the case, a convex utility function is expected [42].

More strict elimination criteria, such as statewise stochastic dominance and first-degree stochastic dominance should, in theory, better fit the utility function of molecular discovery scientists [43, 44]. However, algorithms that rely on higher order elimination criteria often do not converge towards a feasible exhaustive search size. Our main motivation in developing RAD was the numerous successes of ISE [3, 23-26]. ISE eliminates variable values if their propensity of appearance in the worst scoring stochastic solutions is significantly higher than in the top scoring solutions [3]. This is in fact, a “risk averse” method, as a variable value (in our case, rotamer) that appears twice as often in the worst samples than the best could still be, in one particular case, the “global optimum” solution. We hence, deduced that the “risk” element in highly combinatorial problems comes from the feasibility to reach convergence in realistic time and not from the utility function of the discovery goals.

Two main factors influence our preference for a timely, high quality solution ensemble over a single, “global optimum” solution. The first stems from the technical imperfections of scoring methods as reviewed in the introduction. This was also evident in the examples presented in here: In the TLR4 MD2/CD14 solution ensemble, only 2 out of 8 peptides validated were biologically active. Similarly, in the colicin E9 and its cognate/non-cognate partners there were only moderate correlations when repeating the alanine scanning mutation experiments computationally. Nevertheless, RAD provided a very significant enrichment factor when compared to random search. For example, in the high throughput screening for TLR4/MD2 binding molecules only 2 out of 90,000 were discovered [45], representing an enrichment factor of 11,250 for a RAD solution ensemble over random screening. From this it can be assumed that an ensemble of high quality solutions is more likely to yield active molecules than a single “global optimum” solution using current scoring methods. The other factor stems from the ultimate goal of molecular discovery, which in most cases, is reaching molecules that can function at the organism level. The highest affinity “global optimum” is not guaranteed to possess the optimal absorption, distribution, metabolism, excretion and toxicity (ADMET) properties. A diverse solution ensemble is more likely to have at least one molecule with satisfactory ADMET properties than a single solution. We hence believe that at least in the foreseeable future, algorithms that generate solution ensembles, such as ISE or RAD, will have a significant advantage.

CONCLUSION

In this report we demonstrated that for rigid backbone side chain design, RAD can perform on par with ISE at a fraction of the time. This was done by borrowing the “risk/return” heuristic from the field of finance. As in financial portfolios, the RAD approach suffers one major disadvantage: There is a need to accurately define the “risk” function and deviations in risk functions may yield different results. This requires a deep understanding of the problem components, which is not always available. Unlike RAD, ISE only requires a scoring function and the definition of variables with discrete values. This enables ISE to be used on problems without the definition of a problem-specific risk function. The risk function presented here was only focused on binding affinity. We are currently working towards a risk function that additionally incorporates ADMET properties. Another challenge is to attempt to quantify the “noise” portion of current scoring functions. Hence, for example, a rotamer that is calculated to contribute an average of -2 Kcal/mol by a set of scoring functions with small standard deviation between them will be preferred over a rotamer that averages -2 Kcal/mol with a large standard deviation. The identification and usage of new, innovative risk functions can make the risk/return heuristic borrowed from finance, an important tool for molecular design.

CONFLICT OF INTEREST STATEMENT

The authors whose names are listed immediately below report the following details of affiliation or involvement in an organization or entity with a financial or non-financial interest in the subject matter or materials discussed in this manuscript.

Amit Michaeli, Immanuel Lerner, Amiram Goldblum, Anwar Rayan and Alexandra Vardi are affiliated with Pepticom Ltd., a for-profit organization.

REFERENCES

1. Kortemme, T. and Baker, D. 2004, Curr. Opin. Chem. Biol., 8, 91-97.
2. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N. and Teller, A. H. 1953, J. Chem. Phys., 21(6), 1087.

3. Glick, M., Rayan, A. and Goldblum, A. 2002, Proc. Natl. Acad. Sci. USA, 99, 703-708.
4. Pegg, S. C., Haresco, J. J. and Kuntz, I. D. 2001, J. Comput. Aided. Mol. Des., 15, 911-933.
5. Desmet, J., De Maeyer, M., Hazes, B. and Lasters, I. 1992, Nature, 356, 539-542.
6. Handel, T. M. 1998, Computer (Long. Beach. Calif), 8, 471-475.
7. Lewis, S. M. and Kuhlman, B. A. 2011, PLoS One, 6(6), e20872.
8. Gao, M., London, N., Cheng, K., Tamura, R., Jin, J., Schueler-Furman, O. and Yin, H. 2014, Tetrahedron, 21, 7664-7668.
9. Guntas, G., Purbeck, C. and Kuhlman, B. 2010, Proc. Natl. Acad. Sci., 107, 19296-19301.
10. Benjamin Stranges, P. and Kuhlman, B. 2013, Protein Sci., 22, 74-82.
11. London, N. and Schueler-Furman, O. 2008, Structure, 16, 269-279.
12. Markowitz, H. J. 1952, Finance, 7, 77-91.
13. Paydar, Z. and Qureshi, M. E. 2012, Agric. Water Manag., 115, 47-54.
14. Hoekstra, J. 2012, Proc. Natl. Acad. Sci. USA, 109, 6360-6361.
15. Crowe, K. A. and Parker, W. H. 2008, Clim. Change, 89, 355-370.
16. Phillips, P. 2009, Def. Peace Econ., 20, 1-8.
17. Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S. and Richardson, D. C. 2010, Sect. D. Biol. Crystallogr., 66, 12-21.
18. Dunbrack, R. L. and Karplus, M. 1993, J. Mol. Biol., 230, 543-574.
19. Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W. and Kollman, P. A. 1995, J. Am. Chem. Soc., 117, 5179-5197.
20. Still, W. C., Tempczyk, A., Hawley, R. C. and Hendrickson, T. 1990, J. Am. Chem. Soc., 112, 6127-6129.
21. Wesson, L., Eisenberg, D., Wesson, L. and Eisenberg, D. 1992, Protein Sci., 1, 227-235.
22. Bogan, A. A. and Thorn, K. S. 1998, J. Mol. Biol., 280, 1-9.
23. Noy, E., Tabakman, T. and Goldblum, A. 2007, Proteins Struct. Funct. Bioinforma., 68, 702-711.
24. Rayan, A., Senderowitz, H. and Goldblum, A. 2004, J. Mol. Graph. Model., 22, 319-333.
25. Gorelik, B. and Goldblum, A. 2007, Proteins Struct. Funct. Bioinforma., 71, 1373-1386.
26. Rayan, A., Marcus, D. and Goldblum, A. 2010, J. Chem. Inf. Model., 50, 437-445.
27. Froese, D. S., Michaeli, A., Mccorvie, T. J., Krojer, T., Sasi, M., Melaev, E., Goldblum, A., Zatsepin, M., Lossos, A., Álvarez, R., Escribá, P. V., Minassian, B. A., Von Delft, F., Kakhlon, O. and Yue, W. W. 2015, Hum. Mol. Genet., 24, 5667-5676.
28. Kawai, T. and Akira, S. 2010, Nat. Immunol., 11, 373-384.
29. Kim, H. M., Park, B. S., Kim, J. I., Kim, S. E., Lee, J., Oh, S. C., Enkhbayar, P., Matsushima, N., Lee, H., Yoo, O. J. and Lee, J. O. 2007, Cell, 130, 906-917.
30. Kim, J. I., Lee, C. J., Jin, M. S., Lee, C. H., Paik, S. G., Lee, H. and Lee, J. O. 2005, J. Biol. Chem., 280, 11347-11351.
31. Michaeli, A., Lerner, I., Burger-Kentischer, A. and Rupp, S. 2017, Peptide Agonists and Antagonists of TLR4 Activation, WO2017141248 A1.
32. Osborne, M. J., Wallis, R., Leung, K., Williams, G., Lian, L., James, R., Kleanthous, C. and Moore, G. R. 1997, Biochem. J., 831, 823-831.
33. Wong, S. E., Baron, R. and McCammon, J. A. 2008, Biopolymers, 89, 916-920.
34. Meenan, N. A., Sharma, A., Fleishman, S. J., Macdonald, C. J., Morel, B., Boetzel, R., Moore G. R., Baker, D. and Kleanthous, C. 2010, Proc. Natl. Acad. Sci. USA, 107, 10080-10085.
35. Wallis, R., Leung, K. Y., Osborne, M. J., James, R., Moore, G. R. and Kleanthous, C. 1998, Biochemistry, 37, 476-485.
36. Kühlmann, U. C., Pommer, A. J., Moore, G. R., James, R. and Kleanthous, C. 2000, J. Mol. Biol., 301, 1163-1178.
37. Wallis, R., Moore, G. R., James, R. and Kleanthous, C. 1995, Biochemistry, 34, 13743-13750.
38. Kortemme, T. and Baker, D. 2002, Proc. Natl. Acad. Sci. USA, 99, 14116-14121.
39. Rothschild, M. and Stiglitz, J. E. 1971, J. Econ. Theory, 3, 66-84.

40. Giorgi, D. and De Giorgi, E. 2002, J. Bank Financ., 29, 895-926.

41. Rothschild, M. and Stiglitz, J. E. 1970, J. Econ. Theory, 2, 225-243.

42. Von Neumann, J. and Morgenstern, O. 1944, Princet. Univ. Press, 625.

43. Hadar, J. and Russell, W. R. 1969, Am. Econ. Rev., 59, 25-34.

44. Bawa, V. S. 1975, J. Financ. Econ., 2, 95-121.

45. Wang, Y., Su, L., Morin, M. D., Jones, B. T., Whitby, L. R., Surakattula, M. M. R. P., Huang, H., Shi, H., Choi, J. H., Wang, K., Moresco, E. M. Y., Berger, M., Zhan, X., Zhang, H., Boger, D. L. and Beutler, B. 2016, Proc. Natl. Acad. Sci. USA, 113, E884-93.