Fundamental Theory of the Evolution Force: Gene Engineering utilizing Synthetic Evolution Artificial Intelligence

L. K. Davis

***Abstract*—**The effects of the evolution force are observable in nature at all structural levels ranging from small molecular systems to conversely enormous biospheric systems. However, the evolution force and work associated with formation of biological structures has yet to be described mathematically or theoretically. In addressing the conundrum, we consider evolution from a unique perspective and in doing so we introduce the "Fundamental Theory of the Evolution Force: *FTEF*". We utilized synthetic evolution artificial intelligence (SYN-AI) to identify genomic building blocks and to engineer 14-3-3 ζ docking proteins by transforming gene sequences into time-based DNA codes derived from protein hierarchical structural levels. The aforementioned served as templates for random DNA hybridizations and genetic assembly. The application of hierarchical DNA codes allowed us to fast forward evolution, while dampening the effect of point mutations. Natural selection was performed at each hierarchical structural level and mutations screened using Blosum 80 mutation frequency-based algorithms. Notably, SYN-AI engineered a set of three architecturally conserved docking proteins that retained motion and vibrational dynamics of native *Bos taurus* 14-3-3 ζ.

Keywords 14-3-3 docking genes, synthetic protein design, time based DNA codes, writing DNA code from scratch.

I. Introduction

THE evolution force may be described as a compulsion acting at the matter-energy interface that drives molecular diversity while simultaneously promoting conservation of structure and function. The effects of the evolution force are manifested at all levels of life and are responsible for such processes as the formation of genes and gene networks. Herein, we introduce the FTEF and utilize the FTEF to predict formation of GBBs. From our perspective GBBs are short highly conserved sequences formed as evolution artifacts and are principle components of genes. It is not difficult to assert that DNA and protein are matter based computer programs. When viewing genes from the perspective of a computer algorithm GBBs are analogous to fundamental programming blocks. In the current study, we designed a SYN-AI to identify evolution force promoting formation of these programming blocks and to engineer genes by assembly of GBBs.

The FTEF is based on four evolution force identifiers, (i) evolution conservation, (ii) wobble, (iii) DNA binding state, and (iv) periodicity that allow us to compare the magnitude of evolution force associated with DNA crossovers and GBB formation. While, a strong association between cellular function and the evolutionary conservation of DNA and protein sequence has long been recognized [1]-[5], wobble is classically defined as genetic diversity within the third codon with conservation of amino acid sequence [6]-[12]. Herein, we expand wobble's definition to encompass the achievement of genetic diversity with simultaneous conservation of structure, thusly allowing wobble to be quantifiable at all structural levels. We establish DNA binding states as evolution force indicators based on the assumption that the association of energy and life is inseparable and we assert that interaction of evolution force at the matter-energy interface may be characterized by DNA binding states [13]-[16]. There also exists a strong correlation between sequence periodicity and conservation of structure and function as demonstrated by genome Fourier spectrums, thusly, we propose that periodicity is an indicator of evolution force. Prominently, we show herein that the application of these four rudimentary identifiers in conjunction with selection pressure is sufficient to engineer genes de novo.

In order to simulate evolution, SYN-AI integrates a gene-partitioning model that assumes contemporary genes evolved from a single ancestor that expanded to the modern gene pool. Thusly, FTEF is in agreement with the "Universal Ancestor" and LUCA "Last Universal Common Ancestor" models, [20], [21]. We reconstruct DNA exchanges occurring during gene evolution and subsequent point mutations due to speciation by performing gene partitioning. Gene sequences are transformed into DNA secondary (DSEC) and tertiary (DTER) codes in correlation with protein hierarchical structure levels. Thusly, we introduce a time dimension to the DNA code that allows us to fast-forward evolution processes while dampening the effects of point mutations that lead to disruption of protein function. The application of time based DNA codes also allows for conservation of global and local protein architecture as GBBs are conserved from LUCA and have been tested by the evolution process. In terms of hierarchical structure, the DSEC captures evolution on the GBB scale in the range of 19 – 21 base pairs, wherein the DTER captures evolution an order of magnitude higher at the super secondary structure level. Thusly, exchange of genetic information during synthetic evolution is like the swapping of GBBs in a game of Legos and is in agreement with the 'Domain Lego' principle described in [22], [23].

We proved FTEF by proof of concept employing SYN-AI

L.K. Davis was with the Cooperative Agriculture Research Center (CARC) at Prairie View A & M University, Prairie View, TX 77446 USA. He is now at the Gene Evolution Project, LLC (e-mail: lkdavis.geneevolutionproject@gmail.com).

to engineer a set of 14-3-3 ζ docking proteins using the *Bos taurus* 14-3-3 ζ docking gene as an engineering template. GBBs were identified by the magnitude of evolution force associated with DNA crossovers and crossovers simulated by random hybridization of DNA fragments within genomic alphabets comprising the DSEC code. Synthetic super secondary structures were engineered based on the DTER code and constructed by random selection and ligation of GBBs. Following the equilibration of synthetic structure lengths to native structures, we simulated natural selection by applying Blosum 80 mutation frequency and PSIPRED secondary structure based algorithms to select synthetic super secondary structures for gene engineering. 14-3-3 ζ docking genes were engineered by randomly selecting and ligating synthetic structures from appropriate DTER libraries. Notably, SYN-AI constructed a library of 10 million genes that yielded three architecturally conserved docking proteins based upon theoretical closeness of hydrophobic interfaces and active sites to the native *Bos taurus* docking protein. Synthetic 14-3-3 ζ docking protein structure was confirmed by I-TASSER three-dimensional structure estimations and retention of cooperative communication between protein domains confirmed by normal mode analysis.

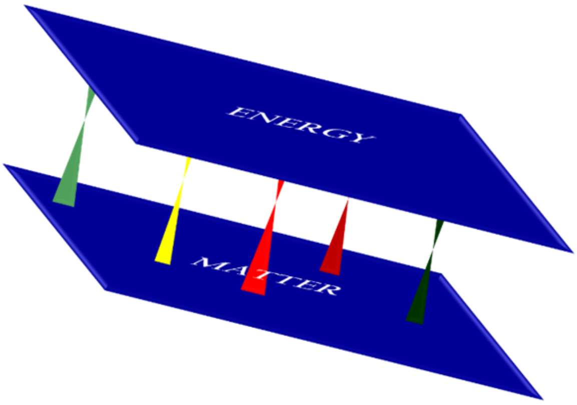

Fig. 1 The Matter-Energy Interface

II. Theory

A. FTEF

We state herein that the evolution force is a compulsion acting at the matter-energy interface that drives genetic diversity while simultaneously conserving biological structure and that the dynamics of the matter-energy interface do not act independently of evolution's tendency toward conservation. We further hypothesize that the four principle identifiers of evolution force are 1) evolution conservation, 2) wobble, 3) DNA binding state, and 4) periodicity.

We established these evolution force identifiers according to the basic engineering format that nature utilizes in respect to genetic relatedness as well as established evolution concepts. To give a simple explanation of FTEF, when considering evolutional conservation of structure we can use the example of bone structure. Human legs comprise of an upper leg having a femur and a lower leg comprising of a tibia and fibula. These structures are conserved in a variety of species in the phylum Chordata; thusly we consider them as artifacts of the evolution force. When considering wobble we look at the aforementioned with respect to the range of genetic diversity covered as they are conserved in genetically distant species. In terms of periodicity, the FTEF hypothesizes that the more often such structures are observed in nature the stronger the influence of the evolution force.

We utilize the FTEF to describe these evolution principals at the molecular level and to engineer genes. However, our theory may be applied to all levels of life. While, the aforementioned are straight forward, the matter-energy interface requires more clarification due to the Theory of Quantum Mechanics and the coexistence of photons as both particles and waves. Not all energy manifest as matter, thusly to describe the effect of energy on gene evolution we took an alternative approach. Whereby, the FTEF views energy and matter as separate but overlapping dimensions that form synapse at critical junctions allowing the sharing of information, Fig. 1. These interfaces are often seen in nature such as the interface of sound waves with the ocular allowing transduction of vibrational energy and its conversion to information by the brain. More ubiquitously, we observe the interaction of photons with photoreceptors allowing conversion of radiation to cellular information. In terms of gene evolution, DNA crossover junctions are a type of matter-energy interface that allow conversion of thermal energy to genetic information. Whereby, enthalpic and entropic factors governed by cellular conditions and sequence contribute to stabilization of the DNA molecule and facilitate the transfer of genetic information. Furthermore, our theory states that evolution force associated with the formation of GBBs may be solved for according to the postulates stated below:

- ***Postulate 1*** - A natural selection system will generate sequences exhibiting positive variation from the mean of a population of randomly evolved sequences occurring during an evolution instance. Whereby, such sequences will display greater evolutionary conservation of the parental sequence.
- ***Postulate 2*** - Due to degeneracy of the genetic code [8], a natural selection system will generate sequences that exhibit higher conservation of protein structure than expected based on mean DNA similarity. The aforementioned is defined as wobble and considered an artifact of the evolution force.
- ***Postulate 3*** - Evolution force regulates molecular diversity at the matter-energy interface in the form of Gibb's free energy dependent DNA base stacking interactions. Thusly, evolution force may be characterized by DNA binding states.
- ***Postulate 4*** - Evolution has a tendency to repeat structures that contribute to survival of an organism where structures that contribute to function occur more frequently. Thusly, evolution force may be solved as a function of sequence periodicity.

B. Evolution Force Identifiers

1) Evolutionary Sequence Conservation

Sequence conservation is strongly correlated with ligand binding, the structure of active sites, protein-protein interaction (PPI) and functional specificity [1]. In a study of DNA binding proteins, it was shown that functionally essential residues are more highly conserved than their counterparts [24]. It has also been established that evolutionary conservation of amino acids contributes to the protein stable core [25]. Relatedly, it has been shown that genes that encode proteins involved in numerous PPI are more evolutionarily conserved than genes encoding less-prolific interactors [26].

The FTEF describes GBB formation as a function of evolutional conservation. Thusly, our theory is in agreement with the Fundamental Theory of Natural Selection as it captures the effects of natural selection on gene evolution by identifying fit haplotypes [27] whereby, the FTEF identifies GBBs based on the magnitude of evolution force about evolution conservation engine ϵ, where ϵ describes conservation at DNA and protein levels and is a function of evolution vectors ϵ^c_{DNA} and ϵ^c_{Pro}. These position vectors reflect DNA crossover homology to the parent sequence in respect to a rigid body comprising full enumeration of DNA crossovers occurring during an evolution instance. They report the position of DNA and protein sequences resulting from DNA crossovers in the evolution potential field and are functions of similarity vectors X_i and X_j that compare recombinant DNA and protein sequences to parental in terms of physiochemical properties, volume, hydrophobicity and charge. The rigid body generates the evolution potential field, wherein relative position of DNA crossovers to the rigid body characterizes their evolutional advantageousness with more distant crossovers being more evolutionally advantageous. Relative positions are described by weighting similarity vectors by evolutional weights W_d and W_p as given in (2) and (3). By applying these weights we normalize the relative position of a sequence in the potential field to all other DNA crossovers within the potential field and the full enumeration of DNA crossovers back to LUCA.

$$\epsilon = \epsilon^c_{DNA} \cdot \epsilon^c_{Pro} \tag{1}$$

$$\epsilon^c_{DNA} = W_d \sum_{i=1}^{GBB} X_i\,,\ \ i = nucleotide \tag{2}$$

$$\epsilon^c_{Pro} = W_p \sum_{j=1}^{GBB} X_j\,,\ \ j = residue \tag{3}$$

Evolution weights W_d and W_p describe the rigid body center of gravity, thusly describe the origin of the evolution potential field. They are functions of recombinant pool mean DNA μ_s^{DNA} and protein μ_s^{Prot}similarity vectors, thusly describe positions of all DNA crossovers in the potential field. Mean similarity is solved by the summation of DNA X_i and protein X_j similarity vectors occurring within sequence space ($sspace^r$) where, $sspace^r$comprises all orthologue-paralogue gene sequences at a selected identity threshold. Evolutional weight is solved in respect to the total number of DNA crossovers (N), thusly reflects full enumeration of DNA crossovers occurring within the evolution potential field.

$$W_d = \frac{1}{n\mu_s^{DNA}}\ and\ \mu_s^{DNA} = \frac{1}{N}\left[\sum_{DNA=1}^{sspace^r}\sum_{i=1}^{GBB} X_i/n\right] \tag{4}$$

$$W_p = \frac{1}{n\mu_s^{Pro}} and\ the\ \mu_s^{Pro} = \frac{1}{N}\left[\sum_{Prot=1}^{sspace^r}\sum_{j=1}^{GBB} X_j/n\right] \tag{5}$$

2) Molecular Wobble

Wobble evolved during expansion of the genetic code from a simple triplet code expressing a few amino acids in which only the middle position was read as proposed by [28] to the modern genetic code comprising 64 codons and 20 amino acids. This is further corroborated by Wu, who suggested evolution of the modern code from an intermediate doublet system, where only the first and second codon positions were read and the third position served as a structural stabilizer [29]. These hypotheses are corroborated by evolution remnants displayed in aminoacyl tRNA synthetases supporting evolution of the modern genetic code from a more primitive ancestor [30]. Moreover, they support the "*Coevolution Theory*" which suggests the genetic code is an imprint of prebiotic pathways that evolved over a 3 billion year period and that were fixed in LUCA [31].

Due to coevolution of wobble with the genome, FTEF views wobble as one of the four principle evolution force identifiers. Wobble allows the evolution force to balance fitness and adaptation by simultaneously conserving protein sequence and introducing genetic diversity in the third codon position. Due to amino acid groupings, mutation in neighboring codon positions also results in genetically close amino acid sequences. Thusly, we define wobble in a more generic fashion allowing us to capture the property in all three codon positions. We solve for wobble ω_m characterizing a DNA crossover by overlapping position vectors ϵ^c_{DNA} and ϵ^c_{Pro} (6), thusly do not discriminate the third codon position. The resulting relationship is a good indicator of evolution force as it is reflective of parallel hierarchical sequence transitions defining multiple molecular states within a sequence space. FTEF designates wobble as a function of genetic displacement x over time t, where t is the number of evolution cycles required to achieve a genetic step of distance x. Displacement of the protein position vector respective to the DNA position vector in the evolution potential field is described by $x = (\epsilon^c_{Pro}/\epsilon^c_{DNA}) - i_n$ where, i_n is an element of identity vector $\hat{\imath}$ that characterizes expected positions of DNA crossovers in the evolution potential field and where $\forall\ i_n = 1$.

$$\omega_m = \frac{x}{t}, where\ x = \frac{\epsilon^c_{Pro}}{\epsilon^c_{DNA}} - i_n,\ \ i_n \in \{\vec{\imath}\}\ \text{and}\ \forall\ i_n = 1 \tag{6}$$

3) DNA Binding States

FTEF assumes synthetic evolution processes simulated by SYN-AI mimic evolution; thusly DNA binding states occurring during simulations are analogous to DNA crossovers occurring during meiosis as supported by previous studies showing the anticipatory effects of DNA shuffling [32].

Genetic diversity occurs by processes such as DNA crossovers and translocations that result in gene duplication, inversion, insertion and deletion [33], [34]. It is widely accepted that these processes result in relaxation of evolutional stringency allowing speciation and random point mutations by neutral evolution [35], [36]. Based on FTEF, DNA crossover junctions facilitating these events are a matter-energy interface by which the evolution force conveys information. Thusly, the three other evolution engines introduced herein derive from and are dependent upon DNA binding states. The effect of the relationship between evolution conservation and sequence homology on DNA hybridization and Gibb's free energy is obvious. Less obviously periodicity or gene frequency also occurs as a result of gene duplication and is directly affected by DNA binding states occurring during DNA crossovers. Likewise, wobble evolved in a similar manner as a result of the convergence of environmental conditions on natural selection and subsequent speciation following DNA exchanges. Inclusion of DNA binding states as an evolution engine allows FTEF to agree nicely with more complex theories describing the coevolution of genes and gene networks [27] whereby, formation of coevolution mechanisms described in Jordan is a consequence of DNA binding states that helped form genomic structural constraints, gene regulatory regions and nodes [27].

DNA binding states express the stoichiometric relationship between DNA crossovers occurring over evolution of the gene, thusly account for thermodynamic contributors described by Gibb's free energy using less costly calculations. Thereby, we can track contributions of the evolution force at the matter-energy interface back to LUCA with less computational cost. According to Davis, DNA binding states p^i are a function of annealing probability $A^V_{L-v,L}$ and DNA binding probability P^i_{Keq} [14], [15] whereby, DNA binding states are a function of volume exclusion at the DNA crossover junction and DNA crossover thermodynamic signatures.

$$P^i = A^V_{L-v\,,L} \cdot P^i_{keq} \tag{7}$$

According to Wetmur, annealing probability $A^V_{L-v\,,L}$ distributes volume exclusion V^α [37] characterizing a DNA hybridization over that of the recombinant pool where, V defines overlap length characterizing a DNA crossover and L defines sequence length. Volume exclusion is a function of the length to volume relationship occurring at the DNA crossover junction, whereby the probability of hybridization decreases beyond a critical volume of the hybridization bubble.

$$A^V_{L-v\,,L} = d_v V^\alpha / \sum d_v V^\alpha \,, where\ \alpha = -\tfrac{1}{2} \tag{8}$$

Thermodynamic contributions to the DNA binding state are solved as a function of the equilibrium constant k_{eq} where, k_{eq} of a DNA crossover is an exponential function of Gibb's free energy. Gibb's free energy of hybridization is solved by summation of $G°\,(i)$ standard free energy changes for the 10 possible Watson-Crick nearest neighbors occurring in a DNA crossover [38] whereby, counterion condensation is accounted for by free energies of initiation $G°\,(init\ w/term\ G\cdot C)$ and $G°\,(init\ w/term\ A\cdot T)$. Likewise, $G°\,(i)$ incorporates an entropic penalty $G°(sym)$ for maintaining C2 symmetry of self-complimentary sequences as described in [38].

$$P^i_{keq} = k_{eq} / \sum k_{eq} \tag{9}$$

$$k_{eq} = exp^{-\frac{\Delta G}{RT}} \tag{9a}$$

$$\Rightarrow exp^{-\frac{\sum_i n_i G°(i) + G°\,(init\ w/term\ G\cdot C) + G°\,(init\ w/term\ A\cdot T) + G°(sym)}{RT}}$$

4) Periodicity

A strong peak at frequency 1/3 is observed in Fourier spectrums of genome coding regions [39], suggesting the presence of selection pressure. Saliently, three base periodicity also allows characterization of species based upon their Fourier spectrum [40]. Sequence periodicity may result from gene duplication and subsequent speciation wherein fit sequences are retained by the genome, thusly reflects natural selection as a result of evolutional fitness. The FTEF considers periodicity as an evolution force identifier and characterizes periodicity P^π as the distribution of GBB frequency f_{ij} over global frequency Z. Variable f_{ij} describes oligonucleotide i and peptide j homolog occurrences within the target gene, and Z is the summation of all DNA crossover frequency within the orthologue/paralogue sequence space back to LUCA. Saliently, the P^π variable compares selectivity of a DNA crossover to adjacent sequences at both the DNA and protein level whereby, sequences displaying high periodicity are reflective of selection pressure by the evolution force.

$$P^\pi = \sum_i^{oligo} \sum_j^{peptide} f_{ij}^{gene} / Z \tag{10}$$
$$where\ \ Z = \sum_{n=1}^{sspace} \sum_i^{oligo} \sum_j^{peptide} \ f_{ij}$$

C. Analyzing Evolution Force Utilizing the Linear Model

The *"Linear Model"* considers evolution force both at the DNA and protein level and ignores transitory effects on mRNA transcripts. GBBs are viewed as particles having high momentum through the evolution potential field, whereby we apply Newton's second law of motion to describe particle momentum as described by $p = mv \Rightarrow \epsilon \cdot \omega_m$. FTEF captures a snapshot of evolution by ascribing an imaginary particle mass to evolution engine ϵ and setting genetic velocity analogous to wobble ω_m, whereby ϵ describes evolution effects on sequence homology and ω_m captures codon mutation as well as remnants of the evolution of the genetic code. Thusly, evolution momentum p reflects change in sequence homology during gene evolution as well as the rate of mutation. By applying Newton's second law, we can also describe genetic acceleration of a DNA crossover thru the potential field as a derivative of mutation rate as described in (12).

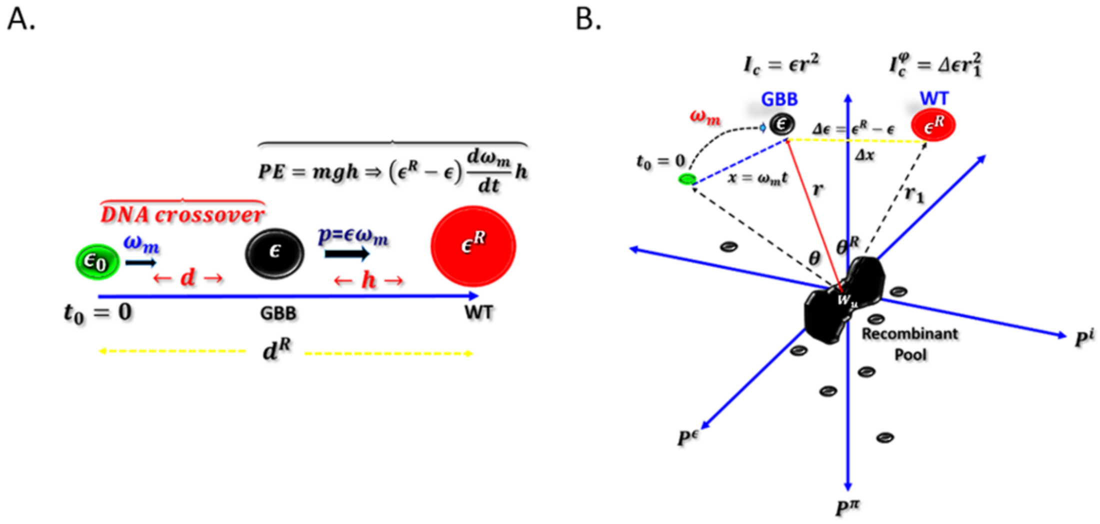

Fig. 2 Linear and Rotation Models

Work performed by the evolution force at the DNA crossover junction can be described by (13):

$$p = mv \Rightarrow \epsilon \cdot \omega_m \tag{11}$$

$$F = \sum ma \Rightarrow \sum \epsilon \cdot \frac{d\omega_m}{dt} \tag{12}$$

$$W = F \cdot d \Rightarrow \sum F \cdot \int (\omega_m + \omega_m^0)\, dt \tag{13}$$

In order to elucidate evolution dynamics, FTEF must describe the relative position of the parent sequence to the rigid body of DNA crossovers formed during evolution of the gene. As the initial position of the parental sequence within the evolution potential field cannot be ascertained we solve for its relative position to the rigid body by viewing it as an ideal DNA crossover characterized by position vectors $\epsilon = 1$ and $\omega_m = 1$. We then describe relative positions (ϵ^R, ω_m^R) of these vectors to the rigid body by applying evolution weight W_d and W_p described in (4) and (5). Momentum p of the WT sequence thru the evolution potential field is a function of evolution vector ϵ^R and relative wobble ω_m^R.

$$p = mv \Rightarrow \epsilon^R \cdot \omega_m^R \tag{14}$$

FTEF solves evolution potential energy (PE) as a steady state, where evolution potential is a function of potential mass m_φ and genetic distance h. Potential mass $m_\varphi = \epsilon^R - \epsilon$ is an imaginary mass characterizing the differential sequence homology remaining between the GBB and parental sequence after DNA recombination and is solved by comparing relative positions of their ϵ vectors in evolution space. Displacement $h = x^R - x$ describes distance of the DNA crossover instance to the WT sequence within the evolution potential field where, x^R is the relative distance of the WT sequence to the initial sequence at t_0 before DNA recombination and x is the distance from the initial sequence to the GBB, Fig. 2 (A). Thusly, PE is a function of differential sequence homology and evolution rate with idyllic DNA crossovers characterized by smaller h values. PE is also solved as a function of evolutional acceleration a_ϵ through the potential field and position vectors (x, y). Vector x describes the time independent rate of change between protein and DNA position vectors (6). Position vector y is the product of vector x and x^R and results from a polynomial derivation of their momentums in the evolution potential field. Our system is pliable as the evolution potential field modifies with each DNA crossover generating fluctuations in sequence homology and genetic velocities. Gene templates having high PE display low evolvability as they are comprised of GBBs characterized by large genetic steps remaining to the WT. The system's kinetic energy (KE) reflects the magnitude of evolution force applied on sequence spaces. GBBs are characterized by high KE, thusly may be defined as sequences displaying high momentum thru the evolution potential field and that exhibit a high degree of evolutional conservation to the WT.

$$KE = \frac{1}{2}\sum \epsilon \cdot \omega_m^2 \Rightarrow \frac{1}{2}\sum a_\epsilon x \tag{15}$$

$$PE = \sum m_\varphi \frac{d\omega_m}{dt} h = \sum a_\epsilon (x^2 - y) \tag{16}$$

$$where\ a_\epsilon = \frac{m_\varphi}{t^2}, y = xx^R\ and\ h = x^R - x$$

The potential energy vector also allows comparison of gene sequence spaces in respect to their mutation rates whereby, the relationship between wobble and incremental potential energy changes ΔPE within a recombinant pool may be described by a first order differential equation $dPE = 2a_\epsilon x dx$ where, the $\overrightarrow{PE}$ characterizing the gene's evolution is described in (17).

$$\overrightarrow{PE} = \sum 2a_\epsilon x dx \tag{17}$$

Total energy TE reflects evolutional advantageousness of the system or gene. It is a function of the evolution force applied on the DNA crossover as well as a function of the relative genetic distance of the DNA crossover to the parent sequence.

$$TE = KE + \overbrace{\sum mgh}^{PE} \Rightarrow \frac{1}{t^2}\sum\left\{\frac{1}{2}\epsilon x^2 + m_\varphi(x^2 - y)\ \right\} \quad (18)$$

The incremental change in the systems total energy ΔTE may also be described by a first order differential equation as described in (19). When considering the effect of wobble with respect to time the relationship may be described as given in (20) where, vector $\overrightarrow{TE}(t)$ describes mutation rate respective to the phylogenetic history of the gene.

$$\overrightarrow{TE} = 2\sum a_\omega\left(\epsilon + 2m_\varphi\right)dx \quad (19)$$

$$\overrightarrow{TE}(t) = -4\sum J_\omega\left(\epsilon + 2m_\varphi\right)dxdt \quad (20)$$

The Lagrangian $\boldsymbol{\mathcal{L}}$ of the system describes the path of the least evolutional resistance. The optimal path for gene formation is enumerated by summation of the differences in KE and PE characterizing DNA crossovers occurring within each genomic alphabet forming the gene's DSEC code. The state $\boldsymbol{S}$ describes the system's evolutional equilibrium and is solved as the integral of $\boldsymbol{\mathcal{L}}$, thusly describes sequence space under the evolution curve with less negative states indicating highly evolvable sequence spaces.

$$\mathcal{L} = \sum KE - PE \Rightarrow a_\epsilon\sum\left\{\frac{1}{2}Ax^2 +\ y\right\} \quad (21)$$

$$\mathcal{S} = \int_{t_0}^{t_f}\mathcal{L}dt \Rightarrow \int_{t_0}^{t_f}\left(a_\epsilon\sum\left\{\frac{1}{2}Ax^2 +\ y\ \right\}\right)dt = -v_\epsilon\sum\left\{\frac{1}{2}Ax^2 +\ y\ \right\}, where\ \ A = \frac{\epsilon}{m_\varphi}\ \ and\ \ v_\epsilon = \frac{m_\varphi}{t} \quad (22)$$

D. Analyzing Evolution Force Utilizing the Rotation Model

The 'Rotation Model' analyzes evolution force associated with GBB formation as a function of evolutional inertia. DNA crossover instances are analogous to particles revolving a rigid body of particles comprised of all DNA crossovers back to LUCA. Evolution force is a function of moments of inertia characterizing the DNA crossover and its acceleration about the respective evolution engine. Evolutional inertia of a particle is described in (23) where, radius r is the standard deviation of evolution vector E from the rigid body, and where vector E is an element of the four evolution engines. Thereby, evolution force τ_E is solved as a measure of central tendency respective to the phylogenetic history of the gene (24). Acceleration about the rigid body is the derivative of mutation rate characterized by wobble vector ω_m, thusly reflects the dynamically changing relationship between DNA and protein positions vectors within the evolution potential field.

Saliently, the Rotation model allows us to solve contributions of each evolution engine to GBB formation by performing multidimensional analysis wherein, we assign each axis of multidimensional evolution space as an evolution engine, see Appendix 'Supplementary Material'.

$$I = Er^2 \quad (23)$$

$$where, E \in \begin{Bmatrix} Evolution\ conservaton, wobble, \\ DNA\ binding\ state, periodicity \end{Bmatrix}$$

$$\tau_E = \sum I \cdot \frac{d\omega_m}{dt} \quad (24)$$

Work applied on the system during the course of evolution is a function of evolutional torque τ_E about the rigid body and radial distance θ where, $\theta = \frac{1}{r}\int(\omega_m +\ \omega_m^0)\,dt$ describes the genetic step of the DNA crossover toward the parent sequence.

$$W = \sum \tau_\epsilon \cdot \theta \quad (25)$$

The system's rotational KE is a function of inertia I_E about evolution engine E and velocity of the particle across the potential field.

$$KE_r = \frac{1}{2}\sum\ I_E \cdot \omega_m^2 \quad (26)$$

The system's total energy TE is the sum of rotational kinetic and potential energies about the rigid body where, rotational potential energy is a function of inertial vector I^φ characterizing potential moments of inertia about evolution engine E and mutation rate. Inertial vector I^φ describes evolution potential in respect to the relationship between recombinant and parental sequence positions in the evolution potential field described by the potential mass vector $m_\varphi = \epsilon^R - \epsilon$ and in respect to all DNA crossovers occurring back to LUCA as characterized by standard deviation r from the rigid body. Thusly, the rotational potential energy describes convergence and divergence of a DNA crossover to the parental sequence in respect to the gene's phylogenetic history and in respect to wobble or mutation rate. State $\boldsymbol{S}$ is a measure of evolutional equilibria characterized by the difference in DNA crossover angular momentum L_{KE} in direction of the KE vector and its angular momentum L_{PE} in direction of the potential energy vector. Reactions characterized by equilibria in the KE direction are more evolutionally favorable. Evolutional states are further a function of position vectors (x, x^R) that describe convergence or divergence of the DNA crossover to the WT sequence and the relative distance of the parent sequence to the rigid body where, $f(x, x^R)$ reflects the hierarchical relationship between protein and DNA position vectors.

$$TE = KE_r +\ PE_r \rightarrow KE_r + \overbrace{\frac{1}{2}\sum I^\varphi \cdot \omega_m^2,}^{PE} \quad (27)$$

$$where, \qquad I^\varphi = m_\varphi r^2$$

$$S = \int_{t_0}^{t_f}\mathcal{L}dt \Rightarrow \int_{t_0}^{t_f}(KE_r - PE_r)dt\ =\ -\frac{1}{2}\sum[L_{KE} - L_{PE}] \cdot f(x, x^R) \quad (28)$$

Incremental changes of the system's TE in respect to mutation rate are described in (29), and changes to equilibria are described in (30):

$$d\overrightarrow{TE} = \sum \tau_E dx, \qquad (29)$$

$$d\vec{S} = -\tfrac{1}{2}\sum L_E\, dx, \qquad (30)$$

E. Evaluating Synthetic Structures

FTEF defines wobble as the conservation of structure in face of genetic diversity. When wobble occurs at the macroscopic level and higher, the tendency is referred to as structural wobble. An example of structural wobble is phyllotaxis, the arrangement of leaves on plants and deformation configurations seen on plant surfaces described in [41]. These Fibonacci-like patterns are conserved across plant species that encompass a broad range of genetic diversity, thusly according to the FTEF they display structural wobble. The FTEF solves for structural wobble as a conditional probability of target structure similarity to the native state where, the probability that a state x_s formed during synthetic evolution will share homology with the native state is a function of closeness probability θ_i, where $i \in P$ comprising physiochemical properties volume, hydrophobicity, charge and folding propensity.

$$wobble = f(x_s|\theta_i),\ where, i \in P \qquad (31)$$

To prevent structural perturbations, SYN-AI performs high-resolution pattern recognition by analyzing discrete sequence spaces occurring across protein structures and walking GBB protein sequences in single steps of one residue. Each step comprises three residues and overlaps the previous step where propensity of characteristic (i) within the sequence space is summated as illustrated in (32). Structural propensity (p) occurring within a discrete sequence space is characterized by the probability density function (δ) as illustrated in (33). Area under the density curve $\int P\, dp$ is normalized by partition function σ describing summation of characteristic i across the structure. The aforementioned allows SYN-AI to characterize the taste of the sequence space. Proteins are characterized by diverse flavors describing discrete changes in physiochemical properties occurring both locally and globally. Closeness of the synthetic structure to the native is described by probability θ_i and solved as a function of synthetic $\delta\,{}_i^{syn}$ and native $\delta\,{}_i^{nat}$ states described in (34).

$$p = \sum_n^{sppace} \sum_i AA\,{}_n^i,\ where\ i \in P \qquad (32)$$

$$\mathrm{\delta} = \frac{1}{\mathrm{\sigma}} \int \mathrm{p\, dp}, \qquad (33)$$

$$\theta_i = 1 - \frac{|\delta\,{}_i^{syn} - \delta\,{}_i^{nat}|}{\delta\,{}_i^{nat}} \qquad (34)$$

In solving the probability of structural state x_s, θ_i is factored across n sequence spaces comprising the structure where, i is an element of S:{ secondary, super secondary and quaternary} structural groups.

$$\prod_{i=1}^{sspace} x_s = \theta_1 \times \theta_2 \cdots\cdots \times\ \theta_{sspace}\ , where\ i \in S \quad (35)$$

FTEF solves wobble as a function of average closeness $\langle Closeness \rangle$ of synthetic and native states whereby, closeness θ_i is summated over n discrete sequence spaces comprising the structure and over characteristic (i) where, N reflects the total number of measurements and i is an element of set P. Wobble occurring across the structure is solved as a function of average structural $\langle Closeness \rangle$ and protein similarity $Prot_s$. Structures exhibiting a greater ratio of closeness to protein similarity display wobble and are characterized by (+) wobble vectors.

$$\langle Closeness \rangle = \frac{1}{N} \sum_i \sum_n [\theta_i]_n\ , where\ i \in P \qquad (36)$$

$$wobble\ = \frac{\langle Closeness \rangle}{Prot_s} - 1 \qquad (37)$$

III. Experimental Methods and Procedures

A. High Performance Computing

SYN-AI experiments were performed utilizing the Stampede 2 supercomputer located at the Texas Advanced Computing Center, University of Texas, Austin, Texas. Experiments were performed in the normal mode utilizing SKX compute nodes comprising 48 cores on two sockets with a processor base frequency of 2.10 GHz and a max turbo frequency of 3.70 GHz. Each SKX node comprises 192 GB RAM at 2.67 GHz with 32 KB L1 data cache per core, 1 MB L2 per core and 33 MB L3 per socket. Each socket can cache up to 57 MB with local storage of 144 /tmp partition on a 200 GB SSD [43].

B. Simulating DNA Crossovers

SYN-AI simulated evolution by partitioning the parental *Bos taurus* 14-3-3 ζ gene into a DSEC and performing 1×10^9 DNA crossovers within genomic alphabets comprising the DSEC. DNA hybridizations were performed at 19, 20 and 21 base pairs allowing us to capture mutations in three open reading frames. DNA hybridization partners were randomly selected across an orthologue/paralogue sequence space constructed by an automated NCBI-Blast. The sequence space comprised of 2.5×10^6 bp of genetic material and genes at a homology threshold of > 80 percent identity to parental *Bos taurus* 14-3-3 ζ. DNA hybridizations were simulated in 3 mM Mg^{2+} and 1.2 mM dNTP at 328.15° kelvin [17]. Gibb's free energy was calculated according to [14] and a penalty assessed for DNA base pair mismatches [18].

C. Simulating Natural Selection

Selection was limited to thermodynamically favored DNA crossovers utilizing an inverse tangent sigmoidal function to scale Gibb's free energy vectors. Free energy vectors were converted to Heaviside nodes by applying an experimental bias and by subsequent transformation utilizing a sinc (x) function in conjunction with a Boolean function. Sequences generating a signal of 1 were considered as GBB candidates and passed thru a cascade of subsequent neural networks. A second round of natural selection utilized pattern recognition filters to remove sequences characterized by long stretches of low sequence homology, thusly lowering the probability of protein perturbations. A third round of selection limited DNA crossovers to those comprised of evolutionarily favored mutations based upon Blosum 80 mutation frequency [19], [20]. In a fourth round of selection, DNA crossovers were limited to those displaying evolution earmarks as characterized by (+) molecular wobble vectors [13]. In a final round of natural selection, DNA crossovers were limited to those characterized by a high magnitude of evolution force.

D. Engineering Synthetic Super Secondary Structures

Parental super secondary structures were identified utilizing STRIDE knowledge based secondary structure algorithms [22], which were utilized to convert the parental 14-3-3 ζ docking protein sequence to a DTER code. A round of natural selection was simulated where neural networks selected against sequences that potentially disrupt protein architecture. Synthetic motifs were then engineered by ligation of GBBs randomly selected from genomic alphabet libraries encompassing 5' to 3' terminals of parental structures. A cleaving algorithm removed 5' and 3' prime overhangs. A second round of natural selection limited selection based on mutation frequency and a final round of selection imposed a secondary structure homology threshold > 88 percent identity to parental 14-3-3 ζ super secondary structures. A standalone version of PSIPRED 4.0 [23] was utilized to evaluate protein secondary structure. Synthetic structures were stored in DTER libraries for writing DNA code.

E. Engineering 14-3-3 ζ Docking Genes

14-3-3 ζ docking genes were engineered by walking the DTER followed by random selection and ligation of synthetic super secondary structures stored in DTER libraries. SYN-AI constructed a library of $1\,X\,10^7$ simulated genes that were passed thru a set of neural networks that evaluated closeness of synthetic protein structural states to native states with a minimal closeness threshold of $>$ 90% identity. Subsequently, selection was limited to proteins characterized by naturally occurring mutations based on BLOSUM80 mutation frequency. A further round of natural selection restricted selection to synthetic 14-3-3 ζ docking proteins having secondary structure identities located within the top quantile of normalized vectors [43]. A final round of selection enriched for functional 14-3-3 ζ docking proteins by comparing synthetic protein active sites and hydrophobic interfaces to that of native *Bos taurus* 14-3-3 ζ wherein, a closeness threshold of > 90% identity was set.

IV. Results and Discussion

A. Analysis of Evolution Force

We validated the FTEF by proof of concept whereby, we utilized SYN-AI to engineer a set of 14-3-3 ζ docking proteins. In order to simulate evolution, a parental *Bos taurus* 14-3-3 ζ docking gene was partitioned into DSEC and DTER codes based on protein hierarchical structural levels. We identified GBBs by performing DNA hybridizations within the DESC and used the DTER as a template for engineering super secondary structures. Evolution force was analyzed utilizing "*Linear*" and "*Rotation*" models. The aforementioned allowed us to simulate probable DNA crossovers going back to LUCA. To enumerate the full range of GBBs, evolution force was evaluated across single and multidimensional planes. Multidimensional analysis is described in the Appendix. SYN-AI engineered 14-3-3 ζ docking proteins by walking the DTER and randomly ligating synthetic super secondary structures stored in DTER libraries. The AI screened for architecturally conserved 14-3-3 ζ docking proteins by simulating natural selection as described in the 'Experimental Methods and Procedures' section.

Linear Model configuration spaces were characterized by broad distributions of evolution force and low resolution of GBBs; however they captured formation of multiple evolution foci suggesting successful simulation of the time-development of the evolution phase space. Sequence 1 was characterized by DNA crossovers distributed around the population expectation at $(\omega_m = 0, \epsilon = 1.0)$ Fig. 1 A (i). However, sequence 2 was characterized by localization of GBB foci in positive and negative evolution phase space indicating presence of strong selection and deselection pressures and a change of biological function, Fig. 3 (ii). Convergence toward WT was signaled by localization of a hotspot at $(\omega_m = 0.45, \epsilon = 1.6)$. Notably, localization of a GBB hotspot in (-) phase space at $(\omega_m = -\ 0.5, \epsilon = 0.3)$ reflects deselection and subsequent speciation that resulted in the change of biological function. Relaxation of evolutional stringency was corroborated by the decrease in evolutional conservation from the population expectation of $\epsilon = 1.0$ to $\epsilon = 0.3$, Fig. 3 A (ii). The pattern of light blue GBB distributions leading to the foci show the time evolution of the sequence block and mutations that lead to the function change. Nonrandom concentric distributions of GBB indicate that the FTEF simulated evolution of the sequence, while concentric yellow hotspots located around foci indicate parallel evolution that resulted in structures of similar function as confirmed by sequence alignments of synthetic 14-3-3 ζ docking proteins performed in [43]. The two less prominent GBB foci localized at $(\omega_m = 0, \epsilon = 0.6)$ and $(\omega_m = -0.1, \epsilon = 1.4)$ indicate the involvement of additional evolution mechanisms. Contrarily, the near normal distribution of DNA crossovers in sequence 1 is due to genetic dispersion resulting from neutral evolution, whereby evolutional noise prevented foci formation.

Contrary to the Linear Model, the Rotation Model achieved high-resolution GBBs denoted by circles, Fig. 3. While SYN-AI used both methods in GBB identification, due to the aforementioned the Rotation Model was predominantly utilized in neural networks. In case of sequence space 1 and 2, we could not capture formation of multiple evolutional foci. However, the model did capture formation of dual foci in adjacent sequence spaces as illustrated in Fig. 4 (A) which shows formation of hotspots in (+) and (-) evolution space. This suggests that our models have varying sensitivity in capturing evolution mechanisms based on sequence space as a result of thermodynamic factors. Formation of GBBs in positive evolution space was characterized by high magnitudes of evolution force and inertia. These results were consistent across higher dimension configuration spaces. Prominently, increased configuration space dimensionality improved GBB resolution, Fig. 4.

Gene sequence spaces exhibited different behaviors due to thermodynamic barriers that form during the evolution process. These barriers formed as a result of the loss of evolutional stringency during speciation, whereby sequences retaining high homology to the ancestor sequence bind more stably in DNA hybridizations due to higher magnitude free energies and lower thermodynamic penalties. We hypothesize that such free energy partitions may have guided evolution processes and are intrinsic components of the evolution force.

We corroborated that the evolution force is a low energy system as work performed in positive and negative directions eliminated each other, Fig. 3 B (iii), (iv). Work $W = \sum \nabla_{x,t}\mathrm{E} \cdot \theta$ performed by the evolution force is a function of the energy gradient $\mathrm{F} = \nabla_{x,t}\mathrm{E}$ and genetic displacement θ; thus, it is a function of the slope of the energy field and dependent on the evolvability of the sequence space. Its dependence on $\nabla_{x,t}\mathrm{E}$ means that it is a superimposition of the resultant of evolution kinetic energy T and potential energy V landscapes, thusly is not static but dynamic. The work configuration space evolves during DNA recombination characterized by changes in genetic distance x. Work associated with GBB formation was distributed around $\theta = 1.0$ which is the expected rotational genetic distance for the population of DNA crossovers. While there was skewing of its mean distribution to (-) phase space, it was counterbalanced by sparse occurrences of sequences in (+) phase space. We expected work to be significantly skewed toward (-) evolution space due to random hybridization of non-homologous DNA sequences. Offset of work in positive and negative phase space suggests that selection protocols implemented by the FTEF are very reliable. Notably, our experiments corroborate findings of [42] and capture a complex interplay of evolution conservation and genetic diversity during gene evolution.

B. Analysis of Synthetic 14-3-3 ζ Docking Proteins

The FTEF was able to capture effects of fitness and evolution rate of discrete protein regions on protein structure as well as provide a description of protein domain formation. Notably, we performed $3\,X\,10^{8}$ DNA crossovers within the 14-3-3 ζ DSEC code and generated $1\,X10^{7}$ docking proteins yet of the three proteins that passed natural selection none contained mutations between residues 99 – 129 and residues 152 – 180 with exception of a $I \rightarrow S$ at residue 106 of SYN-AI-1 ζ and SYN-AI-3 ζ and a $Y \rightarrow S$ at residue 179 of SYN-AI-1 ζ and SYN-AI-2 ζ as we reported in [43]. These regions are almost fully conserved suggesting that they are critical to fitness and comprise very slow evolution rates. When, we superimposed these highly conserved sequence blocks to the SYN-AI-1 ζ three-dimensional structure we observed that all residues were located within the amphipathic groove, Fig. 5.

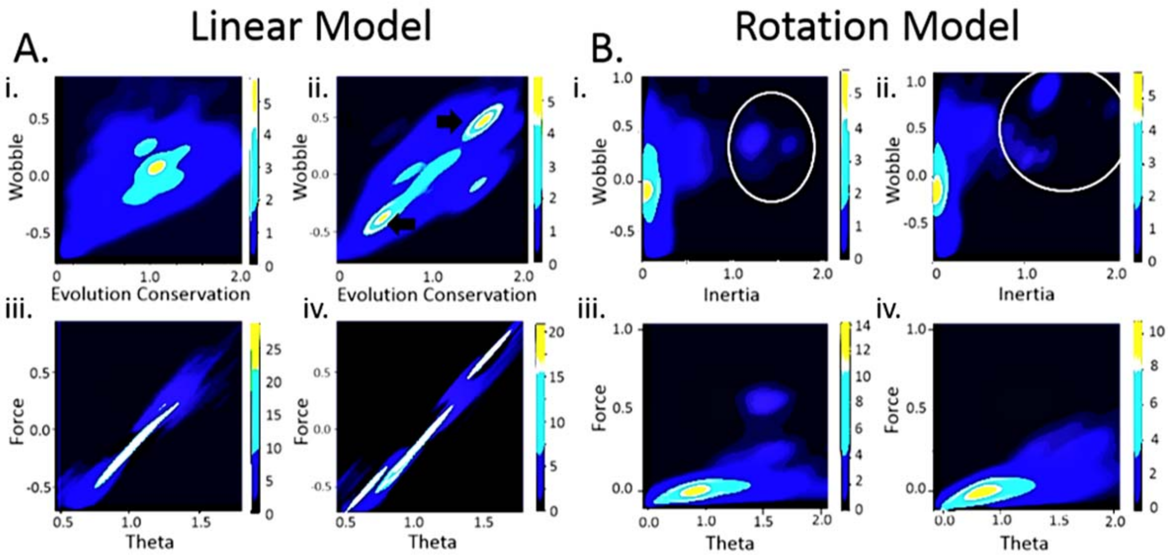

Fig. 3 Evolution Force Linear vs. Rotation Model: Linear Model (A). Evolution force distribution sequence space 1 of the DSEC code (i). Evolution force distribution sequence space 2 (ii). Work distribution sequence space 1 (iii). Work distribution sequence space 2 (iv). Rotation Model (B). Evolution force distribution sequence space 1 (i). Evolution force distribution sequence space 2 (ii). Work distribution sequence space 1 (iii). Work distribution sequence space 2 (iv)

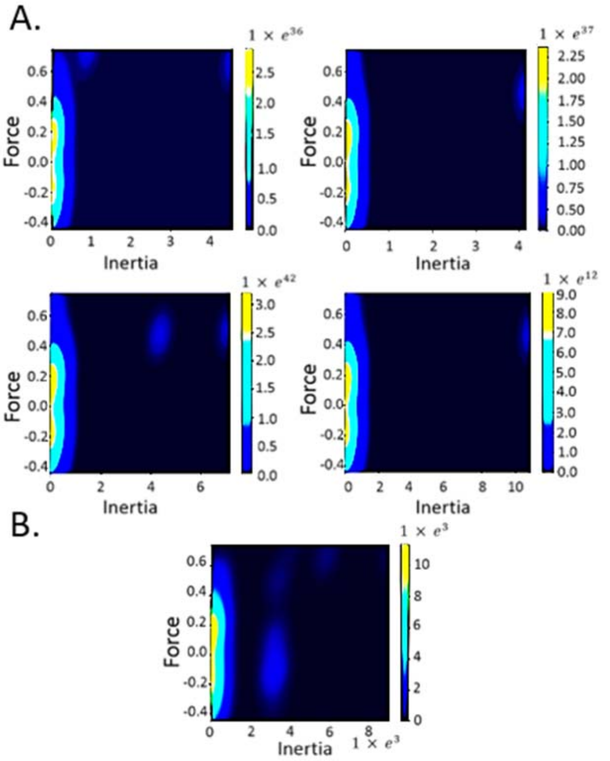

Fig. 4 Evolution Force Distribution across Three and Four-Dimension planes of Evolution. Distribution of evolution force across three-dimension evolution planes is illustrated in GBB density plots (A). Where, we evaluated force distribution across planes, alpha (Top Left), beta (Top Right), gamma (Bottom Left) and rho (Bottom Right). We also analyzed force distribution across a four-dimension evolution plane formed as a resultant of the aforementioned three-dimension planes. Force distribution across the four-dimensional plane of evolution is illustrated in (B). Multidimensional evolution force analysis is described in the ‘Appendix’ section

The amphipathic groove has been reported to be critical to protein function and is the location of the 14-3-3 ζ active site as well as BS01, BS02 and BS03 ligand binding sites [43], [44]. The ability of FTEF to successfully simulate natural selection is corroborated by the positioning of conserved sequence blocks in synthetic 14-3-3 ζ three-dimensional structures. The highly conserved sequence blocks are separated by 23 residues on the protein primary sequence; however, when mapped to the 14-3-3 ζ structure they are located adjacent to each other within the amphipathic groove with overlapping Van der Waals surfaces. The spatial configuration of these sequence blocks suggest that they evolved as separate domains and that in addition to their contribution to the active site and ligand binding they may also play additional functional roles. When, we invert the structure we notice that residues 130 – 151 located between the conserved sequence blocks are associated with the spine of the protein, Fig. 6 (A). As we reported in [45], the spine allows flexibility when performing bend and flex mechanisms during communication between the 14-3-3 ζ active site and C’ terminal helix H3 tail. Although this role is critical to function, our data suggest that this region can tolerate mutation. The highly conserved sequence blocks colocalize to the amphipathic groove as well as play a dual role in protein flexibility allowing the protein to capture the ligand and change configuration to the closed state. We hypothesise this based upon the observed position of these residues when we invert the structure as illustrated in Fig. 6 (A) and by findings we reported in [45]. In addition, ribbon structures depicted in Fig. 6 (B) corroborate that these sequence blocks evolved as separate motifs.

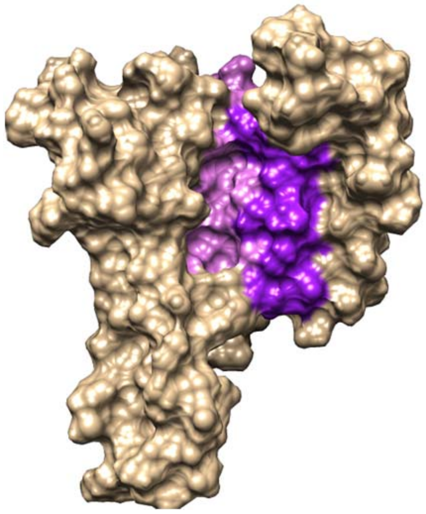

Fig. 5 SYN-AI-1 ζ Structure. SYN-AI-1 ζ three-dimensional structure was estimated using I-Tasser (Zhang Laboratory, University of Michigan). Residues 99 – 129 are colored purple and residues 152 – 180 are colored cyan

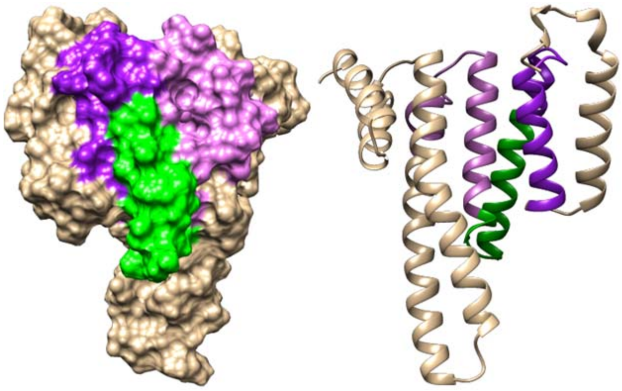

Fig. 6 SYN-AI-1 ζ Structure Reverse View. Residues 99 – 129 (purple), residues 152 – 180 (cyan), and residues 130 – 151 are colored (green). Surface structure (A). Ribbon structure (B)

According to Ghosh, cooperative communications between protein domains is a critical component of protein function [46]. We investigated cooperative communications within simulated 14-3-3 ζ docking proteins using the anisotropic network model, ANM2.1 [47]. We observed that despite a significant sequence divergence of 7.33 achieved by FTEF that based on predicted eigenvalues the global allosteric footprint was conserved as cooperative communications were not lost nor significantly diminished in any region of the docking protein. However, the altered locations and associated eigenvalues of modes suggest altered low frequency vibrations as well as rewiring of cooperate communications within the docking protein, Fig. 7. For instance, near the eigenvalue of 0.1 in the native protein there are two associated modes, however in SYN-AI-1 ζ and SYN-AI-3 ζ there is a 3rd closely associated mode. Likewise, near the 0.7 eigenvalue in the native docking protein we observe two closely associated nodes; however, in SYN-AI-1 ζ and SYN-AI-3 ζ a third mode is introduced to the motion dynamic. There is also an obvious change in cooperative communications involving the three modes located near the eigenvalue of 1.8, whereby the motion of these modes is modified in all three synthetic docking proteins.

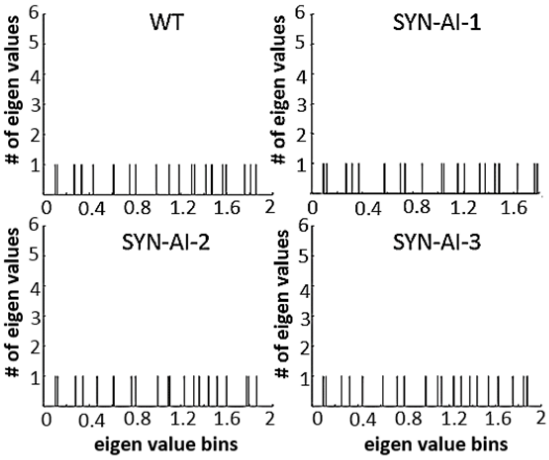

Fig. 7 Normal Mode Analysis. Eigenvalues of the native *Bos taurus* 14-3-3 ζ monomer and synthetic proteins SYN-AI-1 ζ, SYN-AI-2 ζ, and SYN-AI-3 ζ were calculated utilizing the anisotropic network model

We further analyzed synthetic monomers by comparing intra-residue distance fluctuations occurring during normal mode 7, Fig. 8. Native and synthetic distance matrices overlapped well, thusly corroborating that synthetic evolution by FTEF achieved global conservation of 14-3-3 ζ architecture and vibrational dynamics. The ability of FTEF to engineer proteins without disrupting normal modes is critical as 14-3-3 ζ participates in over 230 PPIs and numerous signal transduction pathways. Notably, while the global vibrational footprint was conserved, local distance variations denoted by circled areas suggests that FTEF achieved pathway specific rewiring of cooperative communications.

V. Conclusion

In the current study, we validated FTEF by proof of concept whereby, SYN-AI was utilized to engineer a set of 14-3-3 ζ docking genes utilizing parental *Bos taurus* 14-3-3 ζ as a template for time based DNA codes to guide the engineering process. Gene engineering was simulated by random assembly of GBBs identified by analysis of evolution force. Notably, FTEF resulted in 14-3-3 ζ docking proteins that displayed significant divergence from the parental template while conserving global and local protein architecture. Gene engineering utilizing FTEF also achieved conservation of 14-3-3 ζ normal modes suggesting maintenance of protein vibrational dynamics that regulate signal transduction pathways. We conclude that synthetic evolution by FTEF is an excellent approach for de novo gene engineering.

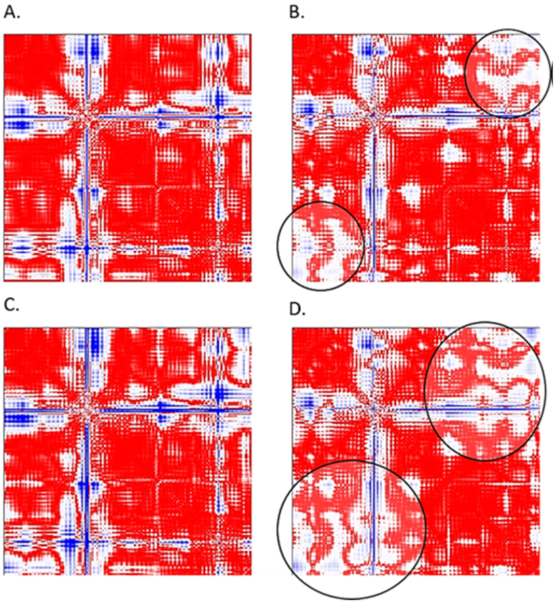

Fig. 8 Distance Matrices. Normal mode 7 vibrational dynamics of native *Bos taurus* 14-3-3 ζ and synthetic docking proteins were evaluated utilizing the anisotropic network model. Native *Bos taurus* 14-3-3 ζ (A), synthetic docking proteins SYN-AI-1 ζ (B), SYN-AI-2 ζ (C), and SYN-AI-3 ζ (D)

Appendix

A. Supplemental Information

1. Evolution Force and Work Distribution in Two-dimension Planes of Evolution

Multidimensional analysis of evolution force is performed using the "Rotation Model" as a function of moments of inertia about selectivity states p_ϵ, p_ω, p_i, and p_π. Selectivity states are calculated by distributing DNA crossover moments of inertia over the summation of inertial moments comprising the rigid body. Inertial moments I_{pE} about evolution engine E are solved by setting the selectivity state analogous to mass and multiplying by variance σ^2_{pE} from the rigid body where, E is an element of the four fundamental evolution engines and the rigid body characterizes full enumeration of DNA crossovers occurring in sequence space ($sspace^r$) as described in (38). Thusly, I_{pE} characterizes moments of inertia about the evolution engine respective to its phylogenetic history back to LUCA. Evolution force $\tau_\epsilon = \sum I \cdot a$ is solved as a function of inertia about the evolution engine and the angular acceleration or the derivative of the mutation rate.

$$I_{p^E} = m \cdot r^2 \Rightarrow \overbrace{\left[I^E / \textstyle\sum_{n=1}^{sspace^r} I^E\right]}^{selectivity} \cdot \sigma_{p^E}^2 \quad (38)$$

$$Where, E \in \{four\ fundamental\ evolution\ engines\}$$

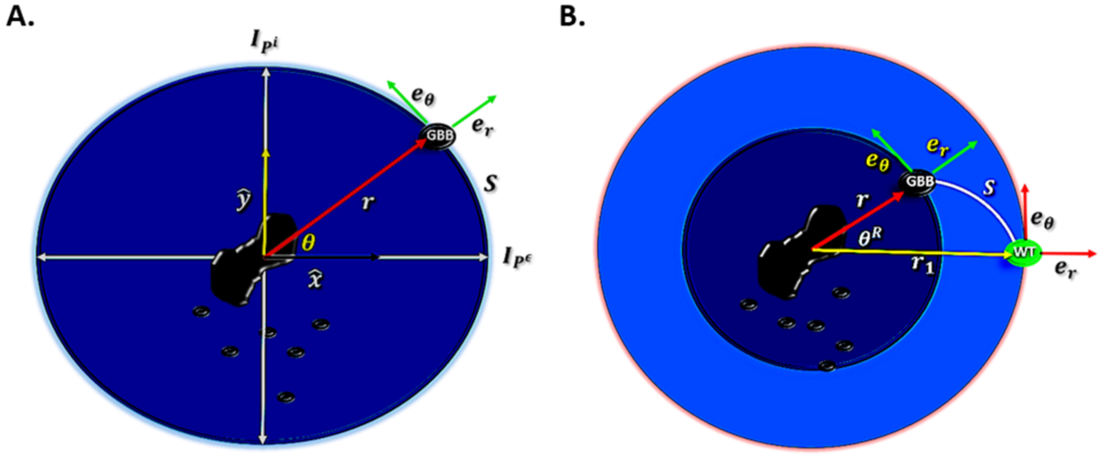

Fig. 9 Identification of GBB Formation and Force Distribution in Two-dimensional Evolution Planes

While, a simple solution of two-dimensional evolution force is given by $\tau_\epsilon = \sum I \cdot a$, we can solve inertia I in $\mathbb{R}^2$ as the resultant of inertial vectors in orthogonal directions with each characterizing an evolution engine, thusly there are six non-redundant inertial vectors formed in $\mathbb{R}^2$ evolution space.

$$I_\alpha = I_{p^i} + I_{p^\epsilon} \quad (39)$$

$$I_\beta = I_{p^i} + I_{p^\omega} \quad (40)$$

$$I_\gamma = I_{p^\omega} + I_{p^\epsilon} \quad (41)$$

$$I_\rho = I_{p^\pi} + I_{p^\omega} \quad (42)$$

$$I_\sigma = I_{p^\pi} + I_{p^\epsilon} \quad (43)$$

$$I_\tau = I_{p^\pi} + I_{p^i} \quad (44)$$

The rotation model also describes evolution force occurring within $\mathbb{R}^2$ as depicted in Fig. 9 where, Evolution force τ_E exerted in formation of a GBB is a function of evolutional torque τ_r applied about the rigid body on fulcrum r as well as angular momentum L_S and torque τ_S of the DNA crossover about displacement vector S as described in (45). S characterizes the genetic step of a DNA crossover toward WT and radius r is the resultant of orthogonal evolution engines.

$$\tau_\epsilon = \sum I \cdot a \Rightarrow \sum(\tau_r + L_S\theta)\hat{e}_r + (\tau_S)\hat{e}_\theta \quad (45)$$

KE is characterized by a polynomial function describing KE distributed about fulcrums r and S. Unit vectors e_θ and e_r describe the evolutional center of gravity as they are normalized expected positions of evolution vectors.

$$KE = \tfrac{1}{2}\sum I \cdot \omega_m^2 \Rightarrow \sum KE_r(\hat{e}_r^2) + 2KE_{rS}(\hat{e}_r\hat{e}_\theta) + KE_S(\hat{e}_\theta^2) \quad (46)$$

We express the system's Lagrangian $\mathcal{L}$ as the difference in two polynomial functions that describe KE about fulcrums r and r_1, Fig. 9 (B) where, radius r_1 describes the relative distance of the parent sequence to the rigid body and radius r is the distance from the DNA crossover to the rigid body. $\mathcal{L}$ is also a function of unit vectors $(\hat{e}_r, \hat{e}_\theta)$ that describe expected linear and rotational evolution distances in respect to the rigid body whereas, the system's State $\mathcal{S}$ is a function of its kinetic and potential states.

$$\mathcal{L} = KE - PE \Rightarrow \tfrac{1}{2}\sum(I\omega_r^2 + 2I\omega_r\omega_S + I\omega_S^2) \cdot f(\hat{e}_r, \hat{e}_\theta) - \tfrac{1}{2}\sum\left(I^\varphi\omega_{r_1}^2 + 2I^\varphi\omega_{r_1}\omega_S + I\omega_S^2\right) \cdot g(\hat{e}_r, \hat{e}_\theta) \quad (47)$$

We express the system's Lagrangian $\mathcal{L}$ as the difference in two polynomial functions that describe kinetic energy about fulcrums r and r_1, Fig. 9 (B). Where, radius r_1 describes the relative distance of the parent sequence to the rigid body and radius r is the distance from the DNA crossover to the rigid body. $\mathcal{L}$ is also a function of unit vectors $(\hat{e}_r, \hat{e}_\theta)$ that describe expected linear and rotational evolution distances in respect to the rigid body.

$$\mathcal{L} = T - V \Rightarrow \tfrac{1}{2}\sum(I\omega_r^2 + 2I\omega_r\omega_S + I\omega_S^2) \cdot f(\hat{e}_r, \hat{e}_\theta) - \tfrac{1}{2}\sum\left(I^\varphi\omega_{r_1}^2 + 2I^\varphi\omega_{r_1}\omega_S + I\omega_S^2\right) \cdot g(\hat{e}_r, \hat{e}_\theta) (48)$$

The motion equation of the evolution configuration space is described by (49), where $\dot{x} \equiv \omega_m$ gives the mutation rate and x is the genetic step.

$$\frac{d}{dt}\frac{\partial\mathcal{L}}{\partial\dot{x}} - \frac{\partial\mathcal{L}}{\partial x} = 0 \quad (49)$$

State $\mathcal{S}$ of the configuration space is a function of its kinetic and potential state and depends on angular acceleration about the evolutional axis.

$$\mathcal{S} = \int_{t_0}^{t_f} \mathcal{L}dt \Rightarrow -\frac{1}{2}\sum \left[\overbrace{\left(r(L_r + 2L_S) + L_S S\right) \cdot f(\hat{e}_r, \hat{e}_\theta)}^{kinetic\ state} - \sum \overbrace{\left(r_1(L_r + 2L_S) + L_S S\right) \cdot f(\hat{e}_r, \hat{e}_\theta)}^{potential\ state} \right] \quad (50)$$

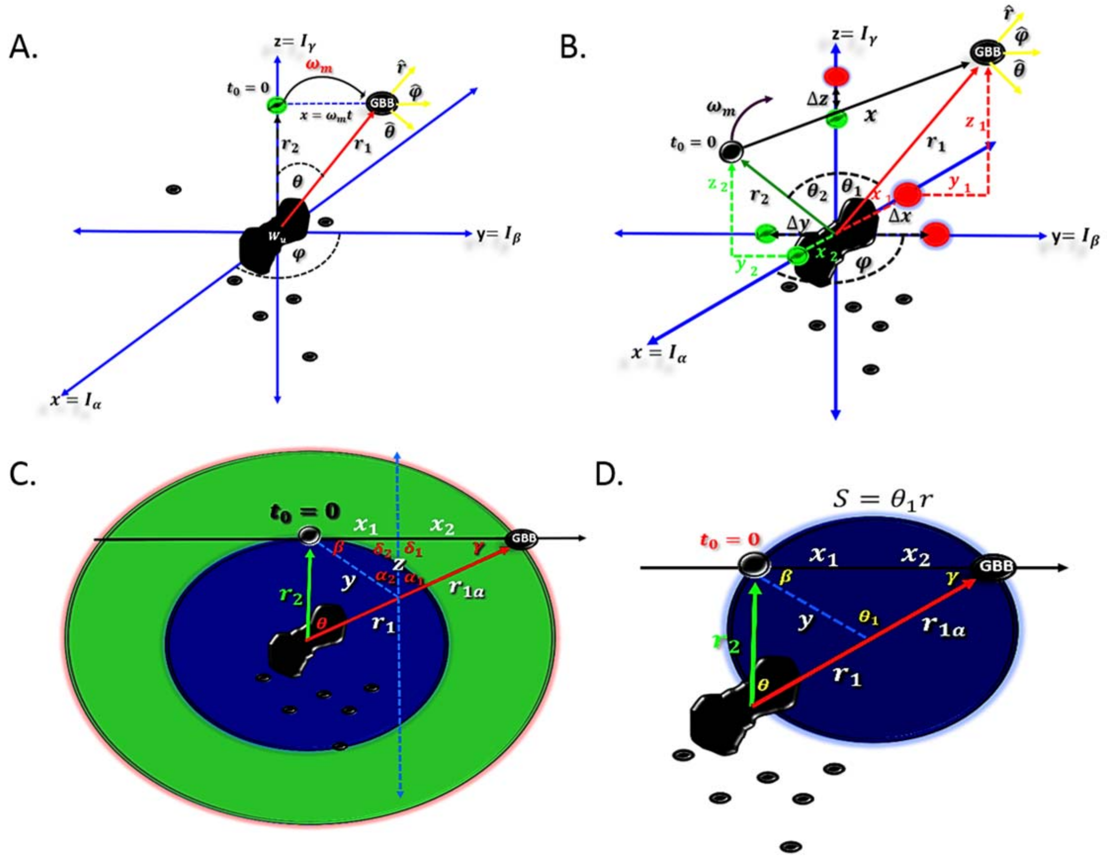

Fig. 10 Inertial distribution in three-dimensional evolution space

2. Evolution Force in Three-Dimension Planes of Evolution

We modeled genomic building block formation in $\mathbb{R}^3$ as a DNA crossover at time $t = 0$ with genetic displacement toward the parental sequence. Where, genetic acceleration in $\mathbb{R}^3$ is a function of molecular wobble, angular displacement θ, and the theoretical azimuth angle φ. Distribution of inertial moments across $\mathbb{R}^3$ phase space is a function of the dot products of inertial vectors I_α, I_β, and I_γ, Fig. 10 (A). Time dependent displacements of the position vector are illustrated in Fig. 10 (B), where distance vector $x = \Delta x + \Delta y + \Delta z$ is the genetic distance between expected and experimental positions of the GBB and represents the incremental inertial change in each direction during a DNA crossover. Unit vectors $\hat{x}$, $\hat{y}$ and $\hat{z}$ depicted (green) give expected DNA crossover positions and are a function of the summation of inertial vectors occurring within the rigid body divided by their magnitude. To solve for arc length S, we modeled expected and experimental positions of genomic building blocks as particles orbiting the rigid body on different evolutional paths. S was solved by rotation about the inertial center and creating a midsection between distance vector x. This permitted re-centering of the inertial center and formation of right triangles that allowed elucidation of angles associated with distance vector x, arc length S, and angle θ_1 utilizing the law of Sines, Fig. 10 (C). We then reset the inertial center and restored the original relationships.

Notably, by expanding force analysis to $\mathbb{R}^3$configuration space, we identified unique genomic building blocks and not only expanded the GBB candidate pool, but also increased the probability of engineering functional genes. When considering distribution of evolution force in three dimensions, thirty non-redundant permutations of inertial vectors form as functions of inertial moments $I_\alpha, I_\beta, I_\gamma, I_\rho, I_\sigma$ and I_τ. They formed both as resultants and dot products of two-dimension configuration spaces. Additionally, three-dimension inertial vectors $I_\alpha, I_\beta, I_\gamma$, and I_ρ form as resultants of vectors I_{Pi}, $I_{P\omega}$, $I_{P\epsilon}$ and $I_{P\pi}$. However, we limit our description to the three-dimension evolution planes described in (49), (50):

$$I_{\alpha} = I_{p^{\epsilon}} + I_{p^{\omega}} + I_{p^{i}} \tag{51}$$

$$I_{\beta} = I_{p^{i}} + I_{p^{\epsilon}} + I_{P^{\pi}} \tag{52}$$

$$I_{\gamma} = I_{p^{i}} + I_{p^{\omega}} + I_{P^{\pi}} \tag{53}$$

$$I_{\rho} = I_{p^{\epsilon}} + I_{p^{\omega}} + I_{P^{\pi}} \tag{54}$$

Four non-redundant permutations of three-dimension planes also form as dot products of inertial vectors characterizing the four evolution engines.

$$I_{\alpha 1} = I_{p^{i}} \cdot I_{p^{\epsilon}} \cdot I_{p^{\omega}} \tag{55}$$

$$I_{\beta 1} = I_{p^{i}} \cdot I_{p^{\epsilon}} \cdot I_{P^{\pi}} \tag{56}$$

$$I_{\gamma 1} = I_{p^{i}} \cdot I_{p^{\omega}} \cdot I_{P^{\pi}} \tag{57}$$

$$I_{\rho 1} = I_{p^{\epsilon}} \cdot I_{p^{\omega}} \cdot I_{P^{\pi}} \tag{58}$$

We describe evolution force in $\mathbb{R}^3$ evolution space as a function of moments of inertia about the four evolution engines as in Fig. 10. Where, the expected position as well as the experimental position of the GBB instance are solved as the resultant of moments of inertia in the $(x, y\ z)$ directions as illustrated in Fig. 10 (B). Distance vector $x = \Delta x + \Delta y + \Delta z$ is the genetic distance between the expected position and the GBB instance and represents the incremental inertial change in each direction during a DNA crossover. Unit vectors $\hat{x}$, $\hat{y}$ and $\hat{z}$ give the expected DNA crossover position and are a function of the summation of inertial vectors occurring within the rigid body divided by their magnitude. It is worth mentioning that angular acceleration in $\mathbb{R}^3$ is solved utilizing the relationship between distance vector x and arc length S whereby, we modelled the expected position vector and GBB instance as particles in $\mathbb{R}^2$ having different radii and paths around the rigid body, Fig. 10 (C). We projected three-dimension evolution space to two-dimensional space by rotating position vectors around the gravitational center to highlight the orthogonal relationship between radii r_2 of the expected position vector and distance vector x. We re-centered the gravitational center by creating a midsection between vector x as illustrated in Fig. 10 (D). The resulting right triangles allowed us to elucidate related angles utilizing the Law of Sines and to solve for the arc length S. The inertial center of mass was then reset to rotational angle θ. The approach enabled us to elucidate evolution dynamics in $\mathbb{R}^3$ by approximating arc length S.

We obtained solutions for evolution system dynamics in $\mathbb{R}^3$ including the evolution force vector $\vec{F}$, evolution force gradient $\vec{\nabla}\vec{F}$ as well as divergence $\vec{\nabla} \cdot \vec{F}$ and curl $\vec{\nabla} \times \vec{F}$ about the rigid body. This allowed analysis of evolutional proneness of genes and gene regions as well as for optimization of experimental conditions. The rotation model describes evolution force about a rigid body of particles characterizing the full enumeration of DNA recombinations over the evolutional history of the gene. The rigid body creates and evolutional gravitational field, whereby as described in [48] the force gradient $\vec{\nabla}\vec{F}$ gives a snapshot of collective directions of acceleration vectors and gravitational force fields. This allows us to analyze directional changes of evolution force within sequence phase spaces and to determine evolution engines that have greater impact on GBB formation under varying thermodynamic conditions. $\vec{\nabla}\vec{F}$ gives a snapshot of phylogenic dynamics of the configuration space, identifying gene regions that are more resistant or susceptible to mutation. Whereby, the dot product of the force gradient and mutation rate $\vec{\nabla}\vec{F} \cdot \omega_m$ gives the rate of change of the evolution force field during time development of the phase space. Divergence $\vec{\nabla} \cdot \vec{F}$ of the evolution force field gives a snapshot of evolution dynamics allowing comparison of configuration spaces by describing the separation of force field lines. We can capture the rate and direction of field expansion and contraction by the expression $\vec{\nabla} \cdot \vec{F} \cdot \omega_m$. Lastly, we evaluate curl of the force vector about the rigid body. This allows for the fine-tuning of experimental conditions by analysis of infinitesimal evolution force field rotations.

$$Let\ r = \left\| I_{\alpha} + I_{\beta} + I_{\gamma} \right\|, \theta = \arctan\left(\left\| I_{\alpha} + I_{\beta} \right\|, I_{\gamma} \right), and\ x = \Delta I_{\alpha} + \Delta I_{\beta} + \Delta I_{\gamma},$$

$$\vec{F} = \vec{I} \cdot \vec{\alpha} \Rightarrow \left[-\hat{r}(F_r + L_S\theta + L_x\varphi^2) + \hat{\theta}\left(F_s - L_x\varphi^2 \tfrac{r_2}{r} \right) + \hat{\varphi}\left(2F_S\varphi \tfrac{r_2}{r} + F_x\varphi \right) \right] \tag{59}$$

$$\vec{\nabla}\vec{F} = F_{\hat{r}} + \varphi\left(2F_S \tfrac{r_2}{r} + F_x \right) \tag{60}$$

$$\vec{\nabla} \cdot \vec{F} = \tfrac{1}{r}(1.5F_r - L_S - L_x\varphi^2)2\hat{r} + \tfrac{1}{x}\left(F_S \tfrac{r_2}{r} + 0.5F_x \right) 2\hat{\varphi} \tag{61}$$

$$\vec{\nabla} \times \vec{F} = \tfrac{\hat{r}}{x}\left[\hat{\varphi}\left(2F_S\varphi \tfrac{r_2}{r} + F_x\varphi \right) + \hat{\theta}(2L_x\varphi)\tfrac{r_2}{r} \right] + \hat{\theta}\left[-\hat{r}(2L_x\varphi)\tfrac{1}{x} + \hat{\theta}(3F_S)\tfrac{1}{r} \right] + \tfrac{\hat{\varphi}}{r}\left[\hat{\theta}(F_S) + \hat{r}(L_S) \right] \tag{62}$$

ACKNOWLEDGMENT

This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562. The authors acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing HPC resources that have contributed to the research results reported within this paper. URL: http://www.tacc.utexas.edu

REFERENCES

[1] Capra J, Singh M, John A (2007) Predicting functionally important residues from sequence conservation. Bioinformatics 23:1875 – 1882.

[2] Capra J, Laskowshi R, Thornton J, Singh M, Funkhouser T (2009) Predicting protein ligand sites by combining sequence conservation and 3D structure. PLoS Comput. Biol. 5(12): e1000585.doc10.1371/journal.pcbi.1000585.

[3] Lawrie D, Petrov D (2014) Comparative population genomics: power and principles for the inference of functionality. Trends Genet. 30(4): 133 – 139.

[4] Ponting C (2017) Biological function in the twilight zone of sequence conversion. BMC Biol. 15 (71). https://doi.org/10.1186/s12915-017-0411-5.

[5] Weinhold N, Sander O, Dominques FS, Sommer L (2008) Local function conservation in sequence and structure space. PLoS Comput. Biol. 4(7). https://doi.org/10.1371/journal.pcbi.1000105

[6] Bulmer M (1987) Coevolution of codon usage and transfer RNA abundance. Nature 325: 728–730.

[7] Bulmer M (1991) The selection-mutation-drift theory of synonymous codon usage. Genetics 129: 897–907.

[8] Crick F (1966) Codon – anticodon pairing: The wobble hypothesis. Mol. Biol. 19: 548 – 555.

[9] Diwan D, Agashe D (2018) Wobbling forth and drifting Back: The evolutionary history and impact of bacterial tRNA modifications. Mol. Biol. Evol. 35 (8): 2046 – 2059. https://doi.org/10.1093/molbev/msy110

[10] Hong Y, Qi L (2011) Mutation and selection on the wobble nucleotide in tRNA anticodons in marine bivalve mitochondrial genomes. PLoS One 6(1):e16147. https://doi.org/10.1371/journal.pone.0016147

[11] Tong KL, Wong JT (2004) Anticodon and wobble evolution. Gene 333:169 – 177.

[12] Xia X (2005) Mutation and selection on the anticodon of tRNA genes in vertebrate mitochondrial genomes. Gene 345: 13 – 20.

[13] Boger D, Fink B, Brunette S, Winston T, Hedrick M (2001) A simple, high-resolution method for establishing DNA binding affinity and sequence selectivity. J. Am. Chem. Soc. 123(25):5878 – 5891.

[14] Davis L (2014). Engineering cellulosic bioreactors by template assisted DNA shuffling and *in vitro* recombination (TADSir). Biosystems 124: 95 – 104.

[15] Moore G, Maranas C, Lutz S, Benkovic J (2000) Predicting crossover generation in DNA shuffling. P. Natl. Acad. Sci. U.S.A. https://www.ncbi.nlm.nih.gov/pubmed/11248060 98(6):3226 – 3231.

[16] Moore G, Maranas C (2002) Predicting out-of-sequence reassembly in DNA shuffling. Theor. Biol. 219: 9 – 17.

[17] Bellesia G, Jewett A, Shea J (2009) Sequence periodicity and secondary structure propensity in model proteins. Protein Sci. 19:141 – 154.

[18] Xiong H, Buckwalter B, Shieh H, Hecht M (1995) Periodicity of polar and nonpolar amino acids is the major determinant of secondary structure in self-assembling oligomeric proteins. P. Natl. Acad. Sci. U.S.A.https://www.ncbi.nlm.nih.gov/pubmed/11248060 92:6349 – 6353.

[19] Leonov H, Arkin I (2005) A periodicity analysis of transmembrane helices. Bioinformatics 21(11):2604 – 2610.

[20] Woese C (1998) The universal ancestor. P. Natl. Acad. Sci. U.S.A. https://www.ncbi.nlm.nih.gov/pubmed/11248060 95:6854 – 6859.

[21] Glansdorf N, Xu Y, Labedan B (2008) The Last Universal Common Ancestor: emergence, constitution and genetic legacy of an elusive forerunner. Biology 3(29). https://doi.org/10.1186/1745-6150-3-29.

[22] Doolittle R (1995) The multiplicity of domains in proteins. Annul. Rev. Biochem. 64: 287 – 314.

[23] Henikoff S, Greene EA, Pietrokovsk S, Bork P, Attwood TK, Hood L (1997) Gene families: the taxonomy of protein paralogs and chimeras. Science 278(5338):609 – 614.

[24] Chang YL, Tsai HK, Kao CY, Hu YJ, Yang JM (2008) Evolutionary conservation of DNA-contact residues in DNA-binding domains. BMC Bioinformatics 9: 53 – 62.

[25] Guharoy M, Chakrabarti P (2005) Conservation and relative importance of residues across protein-protein interfaces. P. Natl. Acad. Sci. U.S.A. https://www.ncbi.nlm.nih.gov/pubmed/11248060 102 (43): 15447 – 15452.

[26] Frywell K (1996) The coevolution of gene family trees. Trends Genet. 12 (9): 364 – 369.

[27] Jordan K, Mariño-Ramírez L, Wolf Y, Koonin E (2004) Conservation and Coevolution in the Scale-Free Human Gene Coexpression Network. Mol. Biol. Evol. 21 (11): 2058 – 2070. https://doi.org/10.1093/molbev/msh222

[28] Crick FHC (1968) The origin of the genetic code. Mol. Biol. 38 (3): 367 – 379.

[29] Wu HL, Elsen J, Bagby S (2005) Evolution of the genetic triplet code via two types of doublet codons. Mol. Evol. 61: 54 – 64.

[30] Fukai S, Nureki O, Sekine S, Shimada A, Vassylyev D, Yokoyama S (2003) Mechanism of molecular interactions for tRNA (Val) recognition by valyl-tRNA synthetase. RNA 9 (1): 100 – 111.

[31] Wong J (1975) A Co-evolution theory of the genetic code. Proc. Natl. Acad. Sci. U.S.A. 72 (5): 1909 – 1912.

[32] Bacher J, Reiss B, Ellington A (2002) Anticipatory evolution and DNA shuffling. Genome Biol. 3 (8): REVIEWS1021. doi:10.1186/gb-2002-3-8-reviews1021.

[33] Schubert I (2007) Chromosome evolution. Curr. Opin. Plant Biol. 10: 109 – 115.

[34] Zhang J (2003) Evolution by gene duplication: an update. Trends Ecol. Evol. 18: 292 – 298.

[35] Hughes A (2002) Adaptive evolution after gene duplication. Trends Genet. 18: 433 – 434.

[36] Lynch M, Katju V (2004) The altered evolutionary trajectories of gene duplicates. Trends Genet. 20: 544 – 549.

[37] Wetmur J, Davidson N (1968) Kinetics of renaturation of DNA. Mol. Biol. 31: 349 – 370.

[38] SantaLucia J (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. P. Natl. Acad. Sci. U.S.A. https://www.ncbi.nlm.nih.gov/pubmed/11248060 95:1460 – 1465.

[39] Wang L, Stein L (2010) Localizing triplet periodicity in DNA and cDNA sequences. BMC Bioinformatics 11: 550 – 557.

[40] Shah K, Krishnamachari A (2012) On the origin of three base periodicity in genomes. Biosystems 107; 142 – 144.

[41] Shipman PD, Newell AC (2004) Phyllotactic patterns on plants. Phys. Rev. Lett. 92(16):168102 – 168101.

[42] Aravind L, Walker R, Koonin EV (1999) Conserved domains in DNA repair proteins and evolution of repair systems. Nucleic Acids Res. 27 (5): 1223 – 1242.

[43] Davis L.(2019) Intelligent design of 14-3-3 docking proteins utilizing Synthetic Evolution Artificial Intelligence (SYN-AI). *ACS Omega 4* (21), 18948-18960. DOI: 10.1021/acsomega.8b03100

[44] Petosa, C.; Masters, S. C.; Bankston, L. A.; Pohli, J.; Wang, B.; Fu, H.; Liddington, R. C. 14-3-3z binds a phosphorylated raf peptide and an unphosphorylated peptide via its conserved amphipathic groove. J. Biol. Chem. 1998, 273, 16305−16310.

[45] Davis L (unpublished) Dancing molecules: Rewiring cooperative communications within 14-3-3 ζ docking proteins. doi: https://doi.org/10.1101/683466

[46] Ghosh A, Vishveshwara S (2008) Variations in clique and community patterns in protein structures during allosteric communication: investigation of dynamically equilibrated structures of methionyl tRNA synthetase complexes. Biochemistry – U.S. 47 (44): 11398 – 11407.

[47] Eyal E, Lum G, Bahar I (2015) The anisotropic Network Model web server at 2015 (ANM 2.0). Bioinformatics 31: 1487 – 1489.

[48] Clawitter CJ. The Ark of Mathematics, Part 5 Vector Calculus Electricity and Magnetism. Clawitter;2013. 38 p.