科学研究費助成事業　　研究成果報告書

平成 ２６ 年　 ６ 月　 ６ 日現在

機関番号：１２６０８
研究種目：基盤研究(B)
研究期間：2011～2013
課題番号：２３３０００８４

研究課題名（和文）ウェットＧＡの分子実現に基づく高度並列型進化計算理論の構築

研究課題名（英文）Massive Parallel Evolutionary Computation Theory based on a Molecular Realization of WetGA

研究代表者
山村　雅幸（Yamamura, Masayuki）

東京工業大学・総合理工学研究科（研究院）・教授

研究者番号：００２２０４４２

交付決定額（研究期間全体）：（直接経費）　　15,700,000円、（間接経費）　　4,710,000円

研究成果の概要（和文）：生命にアイデアを得た進化計算の応用を通じて培われた探索戦略を分子上に実装したウェットGA を提案し、タンパク質tyrRS の基質改変をテストベッドとして、12 世代の理想的な集団分布の推移を実現した。適応度地形の特徴をモデル化し、探索が１つの局所解からはじめられる進化計算の効果的な世代交代モデルに従って実際に進化させた。各世代の変異体のシーケンスを全解読し、変異の蓄積分布を分析した。並列度の高い計算機アーキテクチャーを前提とした新しい進化計算の理論を考案した。

研究成果の概要（英文）：We proposed WetGA implemented on biomoleculer experiments based on the knowledge obtained in the applications of evolutionary computation inspired from real life. We achieved the ideal population distribution transition through twelve generations in the application testbed which engineers tyrRS protein to recognize new substrate. We modeled the difficulty of the fitness landscape then actually evolved a new protein by new generation alternation model which can start with one local minima. We analyzed the mutation accumulation distribution in each generation by reading the sequences of all samples. We also proposed a new evolutionary computation scheme assuming massive parallel computer architectures.

研究分野：総合領域

科研費の分科・細目：情報学　感性情報学・ソフトコンピューティング

キーワード：ウェットGA　進化計算　タンパク質工学　最適化　高度並列GA

１．研究開始当初の背景

生命の「進化」という現象を工学的に利用する試みには、情報科学上の進化計算（遺伝的アルゴリズム：GA)、生命科学上のタンパク質工学（分子進化工学）がある。これらは「変異と選択による最適解の発見」という大雑把な骨格は共有しているが、利用可能な資源・手段の制約や歴史的経緯などによって異なる探索戦略を発達させてきた。

進化計算は初期のナイーブな発想のレベルを抜け、確率論に基礎を置いた信頼性の高い近似最適化手法となった。探索はランダムな初期集団から出発する(a)。変異は複数親の交叉を主に用いる(b)。適応度評価は自由にプログラムできる(c)。一方、直列計算機が想定されており、比較的少ない集団サイズで多くの世代交代を重ねる狭くて深い探索戦略(d)の実装が続けられてきた。

タンパク質工学は生物学実験室ではごく普通の実験技術である。進化計算と対照的に、探索はすでに機能している単独のタンパク質、すなわち局所解から出発する(a)。変異はDNA の実験操作時に自然に導入される。交叉でシャッフルをかけることもあるが、単独親の突然変異が基本である(b)。実験技術として適応度評価は困難(c)であり、定量化できることはめったにない。一方、数多くの変異体を同時並列的に培養することは容易だが、世代交代を重ねることはほとんどなく、広くて浅い探索戦略(d)の実装が普及している。

進化計算における探索の進捗状況は、現世代の個体分布として保持されている。単独親の突然変異よりも、複数親の交叉の方が、個体分布に忠実なサンプリングができる。連続関数最適化では、多峰性の困難な地形において他の近似最適手法より優れた性能を見せるまでになっている。この経験から見ると、タンパク質工学の探索戦略は選択圧が高すぎて局所最適解に初期収束してしまうように見受けられる。実際、タンパク質工学では富士山型の単峰性の地形が想定されており、地球生命の生体分子はこのような地形に沿って進化してきたとまで唱えられている。理論解析でも単独親の突然変異のみの進化戦略（ES）のモデルが利用される。

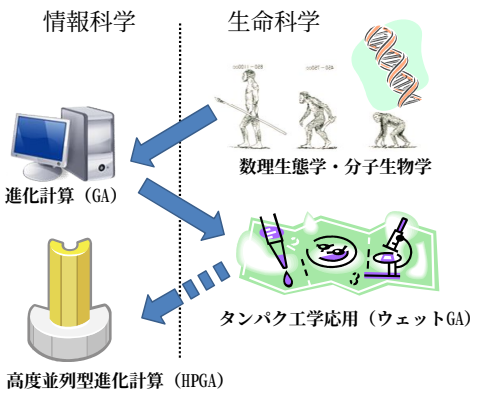

図１ ウェット GA スパイラル

近年の情報科学と生命科学の両分野の技術革新によって、従来の探索戦略を方向づけてきた制約は絶対的条件ではなくなりつつある。これらの暗黙の前提条件の見直しによって、より効率の高い進化計算、より困難な分子進化といった異分野融合スパイラルの次の段階に入る条件が整ったと考える。

２．研究の目的

生命にアイデアを得た進化計算の応用を通じて培われた探索戦略を、分子上に実装したウェット GA を提案し、タンパク質工学に応用し、タンパク質 tyrRS の基質改変をテストベッドとして、12 世代の理想的な集団分布の推移を実現してきた。改変されたタンパク質の活性は実用に耐えるレベルに達している。本研究はこれらの成果を踏まえ、スパイラルの次の段階として計算モデルを洗練化する（図１）。全世代の tyrRS 変異体の遺伝子シーケンスを解読して集団中のエントロピーの推移を計算し、理論との合致を調べる。新しいモデルはネットワークなどを通じて安価に提供されるクラウドなどの並列計算資源に適しており、進化計算の新たな展開を試みる。

着想に至った経緯として、先行する未来開拓「分子コンピューティング」（萩谷)、特定領域「分子プログラミング」（萩谷）では情報科学者と生命科学者の間での緊密な協力関係が不可欠であるとの認識から、分子計算を通じて情報科学・生命科学が同一の概念を共有し、同一の課題に取り組むことを試みてきた。代表者は計画研究において、困難な地形のテストベッドとして選んだチロシル tRNA 合成酵素（tyrRS）（補注１）の基質改変について、理論面では熱力学的遺伝的アルゴリズム（WetTDGA）の枠組みを提案し（補注２)、実験面では目的とする酵素活性を持った分子進化に成功している（補注３)。本研究はこれらの認識と成果の延長上に新たな転回点を見つけようとするものである。

これまでに tyrRS の基質改変の適応度地形の特徴をモデル化し、探索が１つの局所解からはじめられる進化計算の効果的な世代交代モデルを提案して、理論設計に従って実際に tyrRS を進化させた。本研究では、各世代の tyrRS 変異体の遺伝子シーケンスを全解読することによりエントロピーの推移を計算し、モデルを洗練化する。また、特定の世代を選んで tyrRS 変異体を再度進化させ、パラメータ感度の意味でのモデルの正しさを検証する。さらに、これらの知見を発展させて、並列度の高い計算機アーキテクチャーを前提とした新しい進化計算の理論にまとめる。

本研究のように安価に取扱える極めて多数の個体集団を想定した進化計算は新規である。最近のマルチコア技術・ネットワーク技術の進展によって、超高速ではなくても高度に並列化された計算機アーキテクチャー上に、分子に迫る並列度を持った進化計算を

実装できる時代が間近に迫っている。本研究はこのような高度並列型進化計算の理論的基礎として情報科学的に意義がある。

tyrRS の基質改変は方法論のテストベッドであり、ここで得られた方法論上の知見は、そのまま同様の困難なタンパク質の進化に応用可能であることが期待される。タンパク質工学のエキスパートでなくとも、プログラムに与えるデータを入れ替える感覚で、随意のタンパク質を進化させられるのが理想である。なお、改変された tyrRS 分子は、タンパク質への非天然アミノ酸の導入などに発展的に応用可能であり、それ自身有用であることはいうまでもない。

関連研究として、わずかに 1990 年代後半に Woods ら(米)、Baeck ら(蘭)による GA の分子実装があるが、いずれも例示の域を出ていない。分子実装からさらに情報科学にフィードバックさせる試みは例がない。本研究は実用レベルに達した最初の DNA コンピューティングである。

〔補注１〕tyrRS と分子進化

tyrRS はアミノアシル化 tRNA 合成酵素（aaRS）の一種である。aaRS はすべての生物が持っている古いタンパク質で、十分時間をかけて進化してきている。20 種類のアミノ酸と 64 種類の遺伝暗号を対応付けるだけ多種類存在する。酵素としてはほとんど同一の機能を持つにもかかわらず、多種多様の立体構造を持つことが知られていて、広域多峰性の適応度地形を持つと考えられる。最適化の観点から見ると、それぞれの aaRS は「局所解にはまり込んだ」状態にあり、ここを出発点として、大域解あるいは別の遠く離れた局所解を探索する問題は一般には非常に困難である。図２に tyrRS と tRNA の複合体を示す。tRNA のアンチコドン部を正しく認識して物理的にも配列的にも離れた位置でアミノ酸を結合させるため、変異の組合せの影響が、単独の変異の影響からは簡単に決定できない、いわゆるエピスタティスの強い適応度地形を持つ。単峰性であってもエピスタティスの強い地形の探索は困難である。加えて大域的多峰性を持ち、tyrRS は最適化問題としてみたときに困難な対象として、本研究のテストベッドとするにふさわしい。

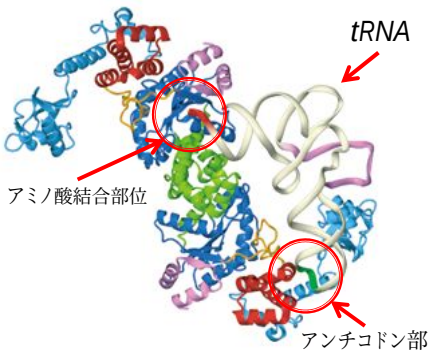

図２高度好熱菌の TyrRS と tRNA の複合体の構造

〔補注２〕WetTDGA

染谷らによる WetTDGA について説明する。タンパク質工学は図３のように最適化問題と対応付けることができるが、利用可能な計算資源の条件が図４のように対照的であるため、計算機上の GA の手順をそのままウェット実装することには意味がない。実験上の制約を反映して設計した WetTDGA 手順を図５に示す。

1. 探索開始点となる個体を増幅した十分な多様性を持つ p 個の個体からなる初期集団 P(0) を生成する。
2. P(t)内の各個体に突然変異をかけ、集団 Pm(t)を生成する。ここで t は世代数を表す。
3. Pm(t)内の個体の交叉により、集団 Pc(t)を生成する。
4. Pc(t)から任意の個体を選択し、選択圧に基づいて受理するかどうか判定する。選択圧は個体の評価値と温度パラメータ T により決定される。この操作を p 回繰り返し、受理された p′個の個体から成る集団 P′(t+1)を生成する。
5. p−p′個の個体を P(t) からランダムに選択し、P′(t+1)に加える。これを集団 P(t+1) とする。
6. 終了条件を満たすならば終了する。そうでなければ、 t に 1 を加え Step 2 に戻る。

Step 4 の選択圧は、焼きなまし法（SA）と同様にメトロポリス法に従って定める。ただし、生物実験では集団内の個々の個体の区別が困難であるため、前世代の集団に含まれる個体の評価値の平均値を SA における状態推移前のエネルギーとみなす。温度パラメータ T は、探索開始時には高温に設定され、探索の進行に伴い徐々に冷却される。これにより、探索序盤では状態推移が比較的容易であることから、集団は分布を広げ図６(a) または(b) のような状態を作り、温度が冷却されるに従い、最適解付近に収束するという挙動が期待される。なお、集団に含まれる個体の評価値の平均値は、一回の培養全体の活性度を測定するなどの方法によって近似値を得ることができる。本手法の特徴には、(i) 各遺伝子の配列の情報を必要としない、(ii) 特定の突然変異オペレータや交叉オペレータに依存しない、(iii) 選択圧を制御することができる、などの点が挙げられる。

最適化	タンパク質工学
目的関数の値(評価値)	望みの機能の活性度
最適解または準最適解	望みの機能をもつタンパク質
解空間(探索空間)	DNA(塩基)またはアミノ酸の配列空間
探索オペレータ	組替えDNA技術

図３ 最適化とタンパク工学の対応関係

	in silico	*in vitro*
遺伝子型の観測性	容易	高コスト
自動化	全自動	部分的
オペレータ	任意	強い制限
親子関係の把握	容易	困難
並列性	数百以下	超並列

図４　in silico, in vitro の条件の違い

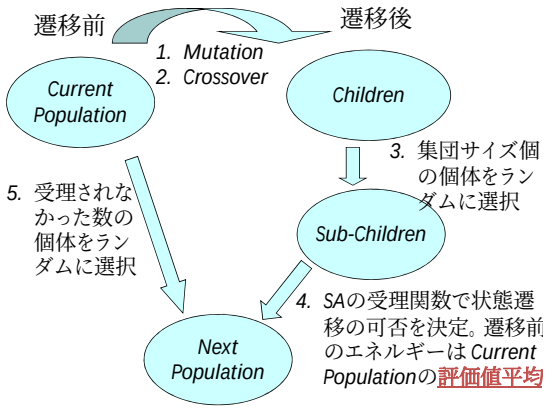

図５　WetTDGA の流れ

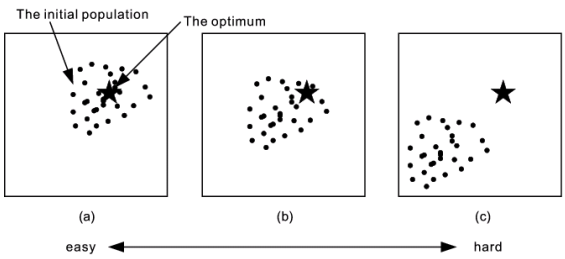

図６　集団分布と進化的探索の容易さ

〔補注３〕tyrRS の基質改変実験

坂本らによる実験について説明する。基質改変は、①STOP コドンを認識して tRNA とアミノ酸を結合する活性の進化と、②他のコドンでは機能しない特異性の進化の２段階からなる。本研究は①までを扱う。②は多目的最適化となる。従来のタンパク質工学とは異なり、初期個体群(a)、交叉(b)、適応度の定量評価(c)、世代交代の実装(d)が求められる。

初期個体群の生成(a)では、タンパク質の進化において、より大域的な探索を目指すには、野生型の遺伝子への突然変異導入を繰り返し行い、集団分布を広げることが要求されるが、変異蓄積によってタンパク質の構造が崩れるなど根本的な形質の破壊が起こる危険性がある。導入されたランダムな変異がタンパク質機能にとって有利にも不利にも働かない中立的な変異だけで構成されていれば、タンパク質は本来の機能を維持し、同時に構造も維持されるであろう。変異導入後も機能を維持している個体だけを残すこととし、これらを初期個体群として用いる。

交叉(b)は集団分布からのサンプラーとして重要な役割を果たす。親遺伝子は、ひとつの野生型に変異を加えて作られたため高い相同性を持つ。PCR による増幅反応で、ポリメラーゼが途中までしか複製を進められないように時間間隔を制御すると１点～多点交叉が実現できる。

適応度の定量評価(c)では、好熱古細菌（M. jannaschii）の チロシル tRNA 合成酵素（MjYRS）の遺伝子と、この酵素によってアミノアシル化されるアンバー・サプレッサー tRNA の遺伝子、およびカナマイシン耐性遺伝子を１つのプラスミド pSup 上にクローニングする。pSup と、アンピシリン耐性と、アンバー・コドン（STOP コドンのひとつ）によって分断された CAT 遺伝子（CAT-Am）を持つプラスミド pTest を大腸菌に導入する。MjYRS が期待される活性を持っているときには、CAT-Am 遺伝子が翻訳されて、MjYRS の活性の大きさに対応して合成される CAT 遺伝子産物の量が増え、高い Cm 耐性を示す。この性質は、様々な濃度の Cm の入った培地を用いれば容易に検定できる。実際の実験では測定誤差を見込んで８段階程度に定量化した。

理論考察で提案した WetTDGA の手順に沿って次のように世代交代させた(d)。

1. ある世代から WetTDGA の選択戦略に従って 100 個のプラスミドを取る（親）
2. tyrRS 遺伝子を交叉 PCR で多様化する（約 1.5 点交叉に調節）
3. カナマイシン耐性遺伝子を持つプラスミドに組み込む
4. 異なる濃度のカナマイシン入りのプレートにまく（適応度評価）
5. コロニーを合計 100 個になるように拾い集めてプラスミド回収（選択）

図８に第 12 世代までの個体分布の変化から抜粋した模式図を示す。活性の測定方法の制約から、上下限を超えたものは合算されている。選択・交叉の繰り返しを通じて、個体分布が次第に高い活性を示す方向にシフトしてゆく様子が見られた。本研究のテーマのひとつは、全世代のシーケンス解読によって、分布の概形の変化だけでなく、エントロピー等によって定量化される遺伝子の多様性の推移を含めて、理論との整合性を検討することにある。

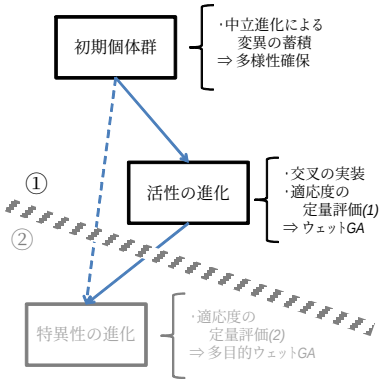

図７　tyrRS 改変実験のフロー

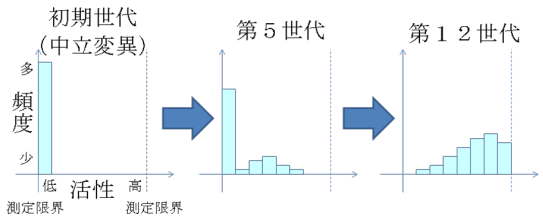

図８　世代交代による個体分布の変化

３．研究の方法

(1) WetTDGA の最適化能力の理論解析(染谷、山村)

①最適化としての困難さを表現する新しいモデルの提案と解析：一般に、野生型遺伝子に加えられた変異がどのようにタンパクの機能に影響を与えるか、という現象のモデル化は困難である。本研究では、個々の現象の精密なモデル化のかわりに最適化の困難さに関わる地形の複雑さをモデル化する。タンパク質工学でよく用いられる NK 地形モデルは、変異間の非加法的相互作用の表現には優れるが、tyrRS の基質改変のような大域的多峰性の困難な最適化問題の地形の表現には不適当である。実験を踏まえてより適切な地形モデルを提案する。

②tyrRS の基質改変実験向け最適パラメータの設計：提案された世代交代モデルから、適応度分布に応じたサンプル個数など、実験向けの最適パラメータを決定する。また、比較用としてネガティブなパラメータ設定も提案し、性能悪化の程度を見積もる。

③提案した地形モデルを tyrRS の基質改変以外にも使えるように一般化する。一般化されたモデルから実験手順やパラメータの見積り等の方法論を確立する。

(2) tyrRS の基質改変のテストベッドを用いた確認実験（山村、小宮、坂本）

①各世代のシーケンス全解読による集団分布の確認実験：坂本らによって、すでに 12 世代分の個体群を得ている。理論解析の正しさとそれに基づく世代交代手順の適切さを証明するために、世代交代によって集団中の変異の分布がどのように変化したのかを、全個体の tyrRS 変異体の遺伝子シーケンスを解読することにより確認する。世代交代を通じて、遺伝子レベルでのエントロピーの変化が、理論解析と合致するかどうかを確認する。

②比較用世代交代バリエーションによる進化実験：成功例が偶然の産物ではないことを証明するために、理論解析から性能の悪化や失敗が予想される手順に基づいて実験した結果についても、同様に世代交代によって集団中の変異の分布がどのように変化したのかを、全個体のシーケンスを解読することにより確認する。

③ここまでに得られた結果を敷衍して、マルチコアやネットワーク上に分散されたクラウドなどの高度な並列性が安価に実現できる計算機アーキテクチャーを前提として、実装コスト・計算量・解の良否の総合的な観点から効果的に実現できる、高度並列型進化計算の枠組みとその性能の理論解析にまとめ上げる。予備的なシミュレーションによって、進化計算研究者の間で常識とされている最適集団サイズの見積りに対して、大きな集団でも性能が維持される感触を得ており、理論化可能と考えられる。

４．研究成果

生命にアイデアを得た進化計算の応用を通じて培われた探索戦略を、分子上に実装したウェット GA を提案し、タンパク質工学に応用してきた。タンパク質 tyrRS の基質改変をテストベッドとして、12 世代の理想的な集団分布の推移を実現した。改変されたタンパク質の活性は実用に耐えるレベルに達している。本研究はこれらの成果を踏まえ、スパイラルの次の段階として計算モデルを洗練化する。全世代の tyrRS 変異体の遺伝子シーケンスを解読して集団中のエントロピーの推移を計算し、理論との合致を調べる。新しいモデルはネットワークなどを通じて安価に提供されるクラウドなどの並列計算資源に適しており、進化計算の新たな展開を目指した。

これまでに tyrRS の基質改変の適応度地形の特徴をモデル化し、探索が１つの局所解からはじめられる進化計算の効果的な世代交代モデルを提案して、理論設計に従って実際に tyrRS を進化させた。各世代の tyrRS 変異体の遺伝子シーケンスを全解読した。各世代の変異の蓄積分布を分析してエントロピーの推移を計算した。また、パラメータ感度の意味でのモデルの正しさを検証するために、tyrRS 変異体を再度進化させる追加実験に向けて世代の選択を試みた。さらに、これらの知見を発展させて、並列度の高い計算機アーキテクチャーを前提とした新しい進化計算の理論を考案した。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　５　件）

1. 染谷　博司、Striking a Mean- and Parent-centric Balance in Real-valued Crossover Operators、IEEE Transactions on Evolutionary Computation, 査読有、17 巻、2013 年、737-754
DOI: http://10.1109/TEVC.2012.2200255

2.石松 愛、畑 敬士、望月 敦史、関根 亮二、山村 雅幸、木賀 大介、General Applicability of Synthetic Gene-Overexpression for Cell-Type Ratio Control via Reprogramming、ACS Synthetic Biology、査読有、2013 年
DOI: http://10.1021/sb400102w

3.関根 亮二、山村 雅幸、Design and Control of Synthetic Biological Systems、Natural Computing and Beyond、査読有、6 巻、2013 年、104-114
DOI: http://10.1007/978-4-431-54394-7_9

4.関根 亮二、山村 雅幸、萩谷 昌巳、木賀 大介、Tunability of the ratio of cell states after the synthetic diversification by the diversity generator、Communicative and Integrative Biology、査読なし、5 巻、2012 年、393-394、
DOI: http://10.4161/cib.20310

5.関根 亮二、木賀 大介、山村 雅幸、Design strategy for an initial state-independent diversity generator、Chem-Bio Informatics Journal、査読有、12 巻、2012 年、39-49
DOI: http://dx.doi.org/10.1273/cbij.12.39

〔学会発表〕（計 2 件）
1.坂本 健作、向井 崇人、大竹 和正、染谷 博司、人為的なコドン再定義から考える遺伝暗号の進化、第 15 回日本 RNA 学会、2013 年 7 月 25 日、愛媛県・県民文化会館・ひめぎんホール

2.林 孝文、山村 雅幸、周波数特性を用いた振動する人工遺伝子回路の自動設計、計測自動制御学会 第 40 回知能システムシンポジウム、2013 年 3 月 15 日、京都工芸繊維大学 松ヶ崎キャンパス

６．研究組織
(1)研究代表者
山村　雅幸（YAMAMURA, Masayuki）
東京工業大学・大学院総合理工学研究科・教授
研究者番号：00220442

(2)研究分担者

(3)連携研究者
坂本　健作（SAKAMOTO, Kensaku）
独立行政法人理化学研究所・チームリーダー
研究者番号：50240685

染谷　博司（SOMEYA, Hiroshi）
東海大学・情報理工学部・講師
研究者番号：00333518

小宮　健（KOMIYA, Ken）
東京工業大学・大学院総合理工学研究科・助教
研究者番号：20396790