

特定領域研究「分子プログラミング」計画研究

平成14年度成果報告書

ナチュラルコンピューティングの分子実現とその設計論

研究代表者 山村雅幸（東京工業大学）

1. 研究目的

遺伝的アルゴリズム(GA), 焼きなまし法(SA)など生命・自然現象に啓発された計算技法をナチュラルコンピューティングと総称する。これらは様々な分野で一定の成功を収めており, 方法論・設計論にも知見の蓄積が見られる。ただし, 生命科学との関わりは一方的なアイデアの借用に留まり, 恣意的な問題設定を離れた普遍的な有効性は明らかではない。

一方, 生命科学ではこれとは独立に生命・自然現象に起源を持つ概念, 例えば焼きなましのようアイデアを日常的に用いており, 変異と選択によって望みの分子を設計する分子進化学は, 確立した分野として成果をあげている。ただし, 情報科学との関わりは, ほとんど同じ現象からアイデアを得ているにもかかわらず, 無交渉なのが現状である。

分子計算には, 情報科学者と生命科学者との間の緊密な協力関係が不可欠であるという大きな特徴がある。分子計算研究は, 類似の概念を異なる意味で用いてきた不思議さを意識する契機ともなった。また, 分子計算の研究対象として, Adleman 流の組合せ問題の超並列厳密解法に加えて, より分子固有の機能を利用する方向を模索しつつある。このように, 分子計算を通じて情報科学・生命科学がついに同一の概念を共有し, 同一の課題に取り組める時期に至ったものと考ええる。こうした相互フィードバックは双方にとって有用なはずである。

本計画研究では, ナチュラルコンピューティングの分子実現を試みる。最適化問題として見たときの適応度地形の特長によって2種類の具体例を用意する。

- (1) 特定の酵素(アミノアシル tRNA 合成酵素(aaRS))の基質特異性の進化的改変を対象とした, 熱力学的遺伝アルゴリズム(TDGA)の分子実現をテストベッドとして生物系・情報系双方向からの検討を加える。aaRS はすべての生物が持っている古いタンパクで, 多種多様の立体構造を持ち, 広域多峰性の適応度地形を持っていると考えられる。TDGA に適した進化実験系を新たに構築し, 実験で得られたデータをもとに, 進化計算, 最適化, 統計物理などナチュラルコンピューティングの側面からの理論解析・シミュレーションを行うことで, 実験条件の改良と同時に理論解析の補強を試みる。
- (2) 伏見らが開発した既存の試験管ウイルスを改良し, ワンポットの試験管中でライフサイクルを回すことのできる人工ウイルスの分子系を構築し, この等温(核酸+タンパク質)増幅系を用いて, 進化リアクターを構築する。この進化リアクター中の試験管ウイルスは, そのリアクター中に設定された環境に適応するように進化するという意味において, 分子人工生命と呼び得る。この進化リアクターを用いて, タンパク質の大域的適応度地形, 並びに, 適応度地形中腹の性質を探查する。タンパクの多くは富士山型の地形をもつことが知られている。

これら地形の性質をタンパク質物性データを基に理論的に調査し、また、その上の適応歩行を考察する。

このような相互フィードバックを通じて、最終的にナチュラルコンピューティングの分子実現のための設計論を構築する。

GA は集団中に蓄積されたビルディングブロック分布のもとで変異・交叉によりサンプル生成する適応的モンテカルロ法である。GA の分子実現による集団的分子計算の計算能力の解明、およびタンパク質というそれ自身強力な機能を持った分子機械の設計は、分子プログラミングの方法および対象として新しい展開である。構築された設計論は既存の分子計算の計算能力や精度の改善にも役立つことが期待される。

実現可能な実験系は計算機上での恣意的な問題設定とは異なる制約を持つ。ナチュラルコンピューティングのシステム論的普遍性にとって、知見の移植性の検討は重要な課題である。本計画研究では GA の忠実な実現ではなく、適応的モンテカルロ法として、あるいは熱力学的最適化法としての設計論の適用を試みる。実験系の手順は狭義の GA には当てはまらないかもしれないが、計算機だけでは得られなかった設計論の補強が期待できる。関連研究に Woods ら(米)、Baeck ら(蘭)のものがあるが、汎用の GA の忠実な実現には適応度評価に無理がある。分子進化を対象とし設計論にまで発展させる試みは例がない。

メンバーの役割分担は次の通りである。

- ・ 全体の取りまとめ
 - 山村雅幸(東京工業大学・大学院総合理工・助教授): 進化計算の理論設計
- ・ 実験サブグループ
 - 坂本健作(東京大学・大学院理学系研究科・助手): 熱力学的遺伝アルゴリズムの実現
 - 伏見 譲(埼玉大学・工学部・教授): 進化リアクターの構築
- ・ 理論サブグループ
 - 樺島祥介(東京工業大学・大学院総合理工・助教授): 統計物理による理論解析
 - 太田元規(東京工業大学・学術国際センター・助教授): 立体構造論からの検討
 - 喜多 一(大学評価・学位授与機構・教授): 最適化の側面からの理論解析

2. 研究成果概要

計画研究の初年度にあたり、主として今後の研究のための情報交換を行った。それぞれの研究成果は次のとおりである。

研究の取りまとめ役として進化計算の理論設計にあたる山村の成果は次のとおりである。

創発システムとは、何もかも人為で作りこむ従来の工学から、システムがボトムアップに生み出す秩序を積極的に利用する新しいシステムの考え方である。遺伝的アルゴリズムに代表される、生命から啓発されたシステムは、そのための強力な要素技術を提供する。遺伝的アルゴリズムにおける交叉というユニークなオペレータの効果について確率過程論を用いた理論解析を行い、進化計算の設計規範として機能分担仮説を提案して、特に関数最適化において定量的な理論設計を可能とした。

実験サブグループで具体的課題(1)に取り組む坂本の成果は次のとおりである。

タンパク質の進化を、複雑な組合せ最適化問題の「分子計算」とみなす立場から研究を行っている。DNA 分子（遺伝子）はタンパク質の機能・構造を符号化した「データ・テープ」とみなされ、タンパク質の発現は「目的関数の計算」プロセスと見なされる。分子計算の並列性をもってしても「最適解」の風潰し的な探索は不可能だが、一方でタンパク質の目的関数が知られていない為、現在の高速な電子コンピュータでもタンパク質の設計には有効ではない。そこで、「進化計算」の計算スキームを取り入れ、コンピュータ・シミュレーションによるドライ実験と、分子生物学のウェット実験の両面からのアプローチを進めている。

ここでは、ウェット実験の方針と成果を報告する。ウェット実験の目的の1つは、タンパク質の現実の適応地形を調査することである。アミノ酸配列の配列空間の大きさを考えると、実際の実験で踏査できる領域は極めてわずかであり、領域踏査の方針は極めて重要になる。従来、野生型（WT）タンパク質の近傍の地形を調べることが行われてきた。しかし、すでに「解」（=WTタンパク質）の得られている空間の近傍を調べるのでは、解探索法としての進化計算の働きを検証することができない。本研究では野生型タンパク質とは明瞭に区別される活性を持つ変異タンパク質を得て、野生型から変異体間の地形を調査するという方針を実地形測定の方針とした。具体的には tRNA に L-チロシン（アミノ酸）を共有結合させる酵素であるチロシル tRNA 合成酵素（TyrRS）を取り上げた。

アミノ酸特異性が WT TyrRS と異なるものを得るために、TyrRS の立体構造を X 線結晶解析法によって決定し、L-チロシン部位の立体構造から改変すべきアミノ酸を 3 残基ずつ 2 セット選んだ。可能なアミノ酸置換は、 $20 \times 20 \times 20 \times 2 = 1,600$ 通りである（遺伝子レベルではこの 3 倍程度の数になる）。遺伝学的な実際の実験によって、16,000 個の変異 TyrRS をスクリーニングし、特異性の変化した変異体を 2 種類得ることができた。今、WT と変異体を含む領域の地形の詳しい調査を進めている。

実験サブグループで具体的課題(2)に取り組む伏見の成果は次の 3 つである。

(1) in vitro Virus 法の改良: 試験管内の遺伝子型表現型対応付け手法である in vitro virus 法は、リボソーム上で新生タンパクを mRNA にリンカーを介して結合させるものであるが、このリンカーを工夫することによりウイルス粒子（mRNA - タンパク結合体）形成効率を改善した。今回のリンカーは、3'末端にピューロマイシンを結合したスペーサーをぶら下げた OMe 核酸オリゴマーが mRNA の 3'末端に Y 字状にハイブリダイゼーションしたものである。この Y 字状の一本鎖部分の 2 つの末端を RNA リガーゼで閉じてヘアピンループにすることにより（Y ライゲーションと呼ぶ）、ピューロマイシンと mRNA が共有結合で結ばれたものが形成される。この方法によれば、mRNA とリンカーは当量でも 100% 近く進行し、形成効率は従来法に比べて遙かに高い。また、in vitro virus のライフサイクルを 1 回回したところ、増殖率も従来法より高いことが確認された。

In vitro virus の増殖段は、従来 RT-PCR 法で行われているが、これを 3SR 等の等温核酸増幅法に転換すれば、自然淘汰型進化リアクターが構築できる。3SR 法は従来 37 の等温過程

で行われてきたが、プライマーのハイブリダイゼーション特異性を上げるために高温で行うことが望ましい。今回は耐熱性 T7RNA ポリメラーゼを用いて 50 の等温過程で増幅する系を構築した。進化リアクターをエミュレートする継代植継ぎ実験にも成功した。ポイントは反応液に 1.4 M トレハロースを投入することである。また、T7 プロモータの 50 の最適配列を RNA-Z 法進化リアクターを用いた in vitro selection で求めた。37 の野生型プロモータとハミング距離 2 だけ離れた配列であった。

(2) タンパク質初期ライブラリーの構築法: in vitro virus のゲノムに載せるべきランダムライブラリーは、終止コドンを含んでいてはならない。また、酸化によるタンパク質変性を避けるためには Met を含んでいない方がよい。あるいは、全てのアミノ酸が等確率で出現するようなライブラリーが望まれる場合もある。このように、コドンの出現頻度を自由に制御出来るような DNA ランダムライブラリーの作成法を 2 種類開発した。上記と同じ Y ライゲーションを利用したブロックシャフリング法 (YLBS 法) と、DNA 合成機を 3 台並列に運転してコンビナトリアル化学のスプリット合成を応用する MLSDS 法である。両者とも、従来法より高品質の初期ランダムペプチドライブラリーを構築できる。

(3) タンパク質配列空間の中立ネットワークの探査: RNA 配列空間上でその存在が確認された中立ネットワークとネットワーク間のインターチェンジが、タンパク質配列空間上の大域的適応度地形としても存在するかどうかの問題となっている。われわれは 10⁸ mer の特定のフォールド 4 種 (型, 型, + 型, / 型) の逆フォールディング問題を太田のポテンシャルを用いて解くことによりこの問題に取り組んだ。各フォールドを与える配列は配列空間をほぼパーコレートしていることがわかり、中立ネットワークは存在するようである。また、全く異なるフォールドの中立ネットワークが互いにハミング距離 5 という近距離まで接近することもわかった。また副産物として、特定のフォールドをするランダム配列球状タンパク集団のエネルギー分布を与える統計学的公式を発見した。これにより、モンテカルロシミュレーションによってエネルギー分布を求める必要がなくなった。

理論サブグループで統計物理による理論解析にあたる樺島の成果は次のとおりである。

情報化社会の進展に伴い様々な分野でデータが大規模化している。そのため、それらを適切に処理する統計モデルへの期待が高まっているが、多くの場合それらの実行は計算論的に困難であるため質の良い近似アルゴリズムの開発が必要となる。1980 年代末、確率推論の研究で「開発」された信念伝播法は大規模な確率モデルの汎用的近似アルゴリズムとして期待されている計算手法の一つである。この手法は多数の要素間の単純な相互作用に基づき状態を反復することで計算を実行するアルゴリズムであり、分子計算など「自然現象を真似た計算に示唆するところも大きい。

従来、信念伝播法は確率モデルにおける要素間の依存関係が木で表現される場合の動作について詳しく調べられてきたが、依存関係に循環経路が存在する場合に関しては相互作用による反復が収束するか否かも含めてその動作が十分には明らかにされていない。本研究では、信念伝播法の動作についての知見を蓄積するために、統計力学で知られているスピングラスモデルに適用し

た場合のダイナミクスを調べた。その結果、巨視的な軌跡の発展規則ならびに微視的な収束性の判定条件を明らかにすることに成功した。

また従来、信念伝播法は確率モデルにおける要素間の依存関係が木や疎結合グラフで表現されている場合の有効性が示されていたが、無数の循環経路が発生する密結合グラフに関してはそれほど良い性能を与えないのではないかと考えられてきた。本研究では、密に結合した確率モデルからの推定問題となる CDMA マルチユーザー復調問題に対して信念伝播法を適用し、従来法を凌ぐ高性能な復調アルゴリズムを開発することに成功した。

理論サブグループで立体構造論からの検討にあたる太田の成果は次の4つである。

(1) タンパク質の立体構造予測サーバの開発：タンパク質の立体構造認識用プログラムの開発を行った。立体構造由来の 3D プロフィールと、配列由来の 1D プロフィールから混合プロフィールを作成し評価した結果、スーパーファミリーレベルの認識能で、1D プロフィールより良い精度を示した。適合度をポテンシャル関数で再評価するプログラムを実装し (PILOT)、自動構造認識コンテスト (CAFASP3) に参加した。結果はそれほど良くはなかったが、今後の課題を得ることができた。

(2) タンパク質の立体構造からの活性部位推定：活性部位の推定は配列のマルチプルアラインメントからの情報抽出に大きくよっている。しかし、配列の進化パターンはかなり複雑で、配列情報のみによる予測には限界がある。そこで、立体構造の情報を機能部位推定に組み入れることを考えた。活性部位は構造安定性からの要請を免除されているので、構造を不安定化する部位によく見られる。つまり、部位置換を導入すると変異体が安定化する可能性が高い。また、タンパク質表面の窪み、例えば穴やクレフトといった、若干内部に埋もれた部位に位置することが多い。上記のような情報を利用して、マルチプルアラインメントから予測される活性部位候補 (保存部位) のうち、構造情報の観点からより確からしい部位を選ぶアルゴリズムを考案した。

(3) ヒト完全長 cDNA のマルチドメインの解析：産総研、生命情報解析センターが 8 月に主催したヒト完全長 cDNA のアノテーションジャンボりに参加し、マルチドメイン構成の解析を行った。ヒトのタンパク質のドメイン構成を大域的に調べたところ、同じドメインが多く繰り返しを持つような、典型的な "動物型" のものが多いことが再確認された。その他、イントロンのフェーズとタンパク質の局在箇所との関係などを調べた。

(4) 小ペプチドのフォールディングシミュレーション：20 残基からなるペプチドのシミュレーションを分子動力学法を使って実行した。今後は軌道の解析などを統計的に実行する。

理論サブグループで最適化の側面からの解析にあたる喜多の成果は次の通りである。

評価値にノイズが加わったシステムの最適化は実験やシミュレーションを通じて行うシステム最適化には必須の技術である。遺伝的アルゴリズム (GA) はこのような用途に適した手法であるが、これまで必ずしも実応用を念頭にアルゴリズムの開発は行われてこなかった。本研究では、評価値にノイズが加わっており、なおかつシステムの評価回数が厳しく制限されている状況下で、探索履歴をメモリに蓄え、未知の評価値を探索履歴を利用して推定する遺伝的アルゴリズムの開発を行ったものである。

3 . 成果リスト

< 山村雅幸 >

- (1) 染谷博司, 山村雅幸, 探索オペレータの機能分担を考慮した進化型計算による関数最適化, 電気学会論文誌 C, Vol.122-C, No.3, pp.363-373 (2002).
- (2) Daisuke Matsuda, Masayuki Yamamura, Cascading Whiplash PCR with a Nicking Enzyme, Proceedings of 8th International Meeting on DNA Based Computers (DNA8), to appear (2002).
- (3) Hiroshi Someya, Masayuki Yamamura, Robust Evolutionary Algorithms with Toroidal Search Space Conversion for Function Optimization, In Proceedings of the Genetic and Evolutionary Computation Conference 2002 (GECCO-2002), 553-560, (2002).
- (4) 山村 雅幸, アクエアスコンピューティング - 生体分子による並列メモリの実現 -, 電気学会誌, Vol.122, No.3, 156-159 (2002).
- (5) Sung-Joon Park, Masayuki Yamamura, FROG (Fitted Rotation and Orientation of protein structure by means of real-coded Genetic algorithm) : asynchronous parallelizing for protein structure-based comparison on the basis of geometrical similarity, Genome Informatics, 13:344-345, 2002
- (6) 山村雅幸, 進化計算の機能分担仮説, 電気学会ナチュラルコンピューテーション共同研究委員会・産業システム情報化研究会資料, 2002

< 坂本健作 >

- (7) Sakamoto, K., Hayashi, A., Sakamoto, A., Kiga, D., Nakayama, H., Soma, A., Kobayashi, T., Kitabatake, M., Takio, T., Saito, K., Shirouzu, M., Hirao, I., Yokoyama, S. " Site-specific incorporation of an unnatural amino acid into proteins in mammalian cells" Nucleic Acids Research 30, 4692-4699 (2002).
- (8) Kiga, D., Sakamoto, K., Kodama, K., Kigawa, T., Matsuda, T., Yabuki, T., Shirouzu, M., Harada, Y., Nakayama, H., Takio, K., Hasegawa, Y., Endo, Y., Hirao, I., Yokoyama, S. " An engineered Escherichia coli tyrosyl-tRNA synthetase for site-specific incorporation of an unnatural amino acid into proteins in eukaryotic translation and its application in a wheat germ cell-free system" Proc. Natl. Acad. Sci. USA, 99, 9715-9720 (2002).
- (9) 坂本健作「DNA コンピュータ」蛋白核酸酵素 (共立出版) Vol. 47, No. 15, pp. 2037-2044, 2002 年

< 伏見譲 >

- (10) Aita T., Hamamatsu N., Nomiya Y., Uchiyama, N., Shibana, Y., Husimi, Y., Surveying a Local Fitness Landscape of a Protein with Epistatic Sites for the study of Directed Evolution, Biopolymers 64, 95-105 (2002)
- (11) Husimi Y., Aita T., Tabuchi I., Correlated Flexible Molecular Coding and Molecular Evolvability, J.Biol.Phys. 28, 499-507 (2002)

- (12) Tabuchi I., Soramoto S., Suzuki M., Nemoto, N., Husimi, Y., An efficient ligation in the making of in vitro virus for in vitro protein evolution, *Biological Proc On-Line* 4, 49-54 (2002)
- (13) Kitamura, K., Kinoshita Y., Narasaki, Nemoto, N., Husimi, Y., Nishigaki, K., Construction of block shuffled libraries of DNA for evolutionary protein engineering: Y-ligation based block shuffling, *Protein Engineering* 15, 843-853 (2002)
- (14) Aita T., Ota M., Husimi Y., An in silico exploration of neutral network in protein sequence space, *J.Theor.Biol.* 221, in press (2003)
- (15) Aita T., Husimi Y., Statistical Formulae for Energy Distribution among a Globular Protein Structure Ensemble, *J.Theor.Biol.* 220, 107-121 (2003)
- (16) Husimi, Y., Evolutionary molecular engineering and the evolvability of biopolymers, 7th International Symposium on Bio-Nanoelectronics: Creation of Nano-Micro Structures in Bioscience & Technology, 7, 14-28 (2002)
- (17) Aita, T., Husimi, Y., Acceleration of Encoded Protein Evolution: in vitro virus, Analysis of local fitness landscapes and Adaptive jump, Santa Fe Institute Workshop on Robustness and Evolvability of Molecules and Microbes, 10 (2002)
- (18) Aita, T., Husimi, Y., Formulation of the energy distribution among a globular protein structure ensemble for the high-throughput screening of sequences, *Genome Informatics* 13, 346-347 (2002)
- (19) Soramoto, S., Ueno, S., Tabuchi, I., Husimi, Y., Design of Various High Quality Random Libraries for in vitro Protein Evolution, *Genome Informatics*, 13, 527-528 (2002)
- (20) 伏見 謙編著, シリーズ・ニューバイオフィジックス 第8巻「生命の起源と分子進化の物理学」(共立出版, 2002)
- (21) Tabuchi I., Soramoto S., Ueno S., Husimi Y., Multi-Line Split DNA Synthesis: a combinatorial method to make a high quality peptide library, *Nucleic Acids Res.* submitted.
- < 樺島祥介 >
- (22) A statistical-mechanical approach to CDMA multiuser detection: propagating beliefs in a densely connected graph Yoshiyuki Kabashima, cond-mat/0210535 (2002)
- (23) Propagating beliefs in spin glass models Yoshiyuki Kabashima, cond-mat/0211500 (2002)
- < 太田元規 >
- (24) M. Ota, K. Kinoshita and K. Nishikawa, Prediction of catalytic residues in enzymes based on known tertiary structure, stability profile, and sequence conservation, *J. Mol. Biol.* in press
- (25) T. Aita, M. Ota and Y. Husimi, An in silico exploration of the neutral network in protein

sequence space, *J. Theor. Biol.* in press

- (26) Y. Isogai, M. Ota, A. Ishii, M. Ishida and K. Nishikawa, Identification of amino acids involved in protein structural uniqueness: Implication for de novo protein design, *Protein Eng.* 15 (2002) 555-560
- (27) K. Homma, S. Fukuchi, T. Kawabata, M. Ota and K. Nishikawa, A systematic investigation identifies a significant number of probable pseudogenes in the *Escherichia coli* genome, *Gene* 294 (2002) 25-33
- (28) T. Kawabata, S. Fukuchi, K. Homma, M. Ota, J. Araki, T. Ito, N. Ichiyoshi and K. Nishikawa, GTOP: a database of protein structures predicted from genome sequences, *Nucleic Acids Res.* 30 (2002) 294-298.
- (29) 太田元規 「あなたにも役立つバイオインフォマティクス」菅原秀明編 共立出版 (2002) 第9章 アミノ酸配列から2次構造を予測する (p69-76) 第10章 アミノ酸配列から立体構造を予測する (p77-84)
- (30) 若山守, 森口充瞭, 太田元規, 西川建 タンパク質の立体構造予測 結晶化しない酵素への応用 *化学と生物* 40 (2002) 452-459

<喜多一>

- (31) 佐野泰仁, 喜多一: 探索履歴を利用した遺伝的アルゴリズムによる不確実関数の最適化, *電気学会論文誌*, Vol. 122-C, No.6, pp. 1001-1008 (2002).
- (32) Yasuhito Sano, Hajime Kita: Optimization of Noisy Fitness Functions by means of Genetic Algorithms using History of Search with Test of Estimation, *Proc. CEC2002*, pp. 360-365 (2002)

4. 代表的な論文

以下の3つはそれぞれ理論, 具体的課題(1), (2)に対応する成果である. その別刷を添付する.

- (1) 山村雅幸, 進化計算の機能分担仮説, *電気学会ナチュラルコンピューテーション共同研究委員会・産業システム情報化研究会資料*, 2002.
- (2) Kiga, D., Sakamoto, K., Kodama, K., Kigawa, T., Matsuda, T., Yabuki, T., Shirouzu, M., Harada, Y., Nakayama, H., Takio, K., Hasegawa, Y., Endo, Y., Hirao, I., Yokoyama, S. "An engineered *Escherichia coli* tyrosyl-tRNA synthetase for site-specific incorporation of an unnatural amino acid into proteins in eukaryotic translation and its application in a wheat germ cell-free system" *Proc. Natl. Acad. Sci. USA*, 99, 9715-9720 (2002).
- (3) Tabuchi I., Soramoto S., Suzuki M., Nemoto, N., Husimi, Y., An efficient ligation in the making of in vitro virus for in vitro protein evolution, *Biological Proc On-Line* 4, 49-54 (2002)

以上