

今月の活動
小椋 信弥

1 今月行ったこと

1.1 PSA/GAc の実装

今月は、これまで主に用いてきた最適化手法である、遺伝的交叉を用いた並列シミュレーテッドアニーリング (PSA/GAc) の実装を行った。従来の PSA/GAc は Fortran で実装していたが、今回モジュールプログラミングの学習を兼ねて、C 言語による実装を行った。C 言語によって実装した PSA/GAc の主な仕様は、以下の通りである。

- 従来は、パラメータを記述したファイルを別途用意する必要があったが、これを ga2k のように実行する際の引数として指定することができる。
- 従来は、対象とするタンパク質ごとに異なる main function を用意する必要があったが、これを一つにまとめ、対象問題はパラメータとして指定するようにした。対象とできるタンパク質は、Met-enkephalin, (Ala)₁₀, (Gly)₁₀, (Val)₁₀, C-peptide, PTH(1-34), BPTI, Protein G の 8 種類である。
- 一回の交叉で複数の子個体を生成することが可能。子個体数はパラメータとして指定することができる。

1.2 文献調査

修士論文で、データベースを用いて β シートを探索するためのシステム構築を行う予定であるため、今回、バイオフィォマティクス~ゲノム配列から機能解析へ¹⁾を読み、データベースを用いたタンパク質立体構造予測の手法についての調査を行った。

多くの場合、データベースを用いたタンパク質の立体構造予測においては、構造を調べたい未知のアミノ酸配列と、立体構造が既知のアミノ酸配列データベースとを比較することによって行われる。すなわち、実験的に解明されたタンパク質が増えてくるに従い、類似の折りたたみを持つ構造を見つけるためにデータベースが用いられるようになった。配列の類似性検索を行う代表的なプログラムとして、FASTA と BLAST がある。これらはいずれも、2 つの配列間でできるだけよいアラインメント、すなわち最適なアラインメントを得るためのツールである。アラインメントは、Fig. 1 に示したように、配列を連続した線の上に書き、文字が一致する部分を同じ列に置き、文字が一致しない部分は不一致として同じ列に

置くか、ギャップ (空白) を置いて次に挿入 (もう一方の配列にとっては欠失) するかして得られる。最適なアラインメントをみつけるために、一致、挿入、欠失のすべての可能性を考慮していくことは非常に困難であり、例えば 300 アミノ酸からなるタンパク質では、 10^{88} 回の比較を行わなければならない。

```

4   4   :   ü   ü   6   7   8   I   <   :
      ||  ||           ||  ||           ||  ||
5   4   :   8   L   6   7   ü   <   <   :
```

Fig. 1 2 つの配列間のアラインメントで、一致、不一致、ギャップ (-) を示す。最良の、すなわち最適なアラインメントは、これらの 3 通りの変化すべて考慮したものである。

1.3 CEC の原稿作成

12 月 8 日から 12 日まで、オーストラリアのキャンベラで開催される 2003 Congress on Evolutionary Computation に投稿した論文が通ったため、現在カメラレディ原稿の作成を行っている。論文の前半 (PSA/GAc の提案と、PSA/GAc によるタンパク質立体構造予測の章) が、これまでに発表した内容と同じであるという査読結果だったため、これを大幅に書き換える予定でいる。

具体的には、PSA/GAc の提案およびアルゴリズムの説明、PSA/GAc によるタンパク質立体構造予測の実験結果を削除し、提案するローカルサーチアルゴリズムを検証するために作成したテスト関数について記述することとした。

2 今後の予定

- CEC の準備
- データベースを用いて β シートを探索する

参考文献

- 1) バイオフィォマティクス ゲノム配列から機能解析へ。David W. Mount, 岡崎康司訳。メディカル・サイエンス・インターナショナル。2002 年 5 月。