頻出パターン解析による重複遺伝子の同定手法

理学専攻　情報科学コース　小中 佳子　(指導教員：瀬々 潤)

1　はじめに

生物が新規機能を獲得する上で，遺伝子重複は重要なイベントのひとつである．重複した一方の遺伝子に変化が起きても機能損失なく，変異を蓄積していくことが出来るからである (図 1)．その一方で，遺伝子の重複を見つけることは必ずしも容易ではない．特に進化的に最近起こった重複の場合，変異の入る量が少なく，複数の領域から採取されたものなのか，単一の領域から採取されたが，配列決定時のミスにより，複数種類の配列が存在するように見えてしまうのかの区別を付けることが必ずしも容易ではないためである．このため，重複遺伝子の探索は，全ゲノム配列を決定し，その後，遺伝子間の配列相同性を調べる手法が多かった．しかし，ゲノム配列の決定は非常にコストが高く容易に決定できるものではなく，特に真核生物では限定的な種でしかゲノム配列が決定できていない．

本研究では，この全ゲノム配列決定を避け，遺伝子領域のみのシークエンスから遺伝子重複が起こっていることを推定する．より具体的には，次世代シークエンサで読まれた大量のリード配列に対し，近縁種のゲノム配列を参照する事で，リードと近縁種ゲノムの間にある変異位置を同定する．そして，その変異位置が共起するリードを調べてグループ化する事で，重複遺伝子の同定を行う．

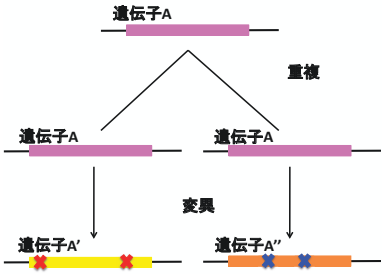

図 1: 遺伝子重複

2　手法

本研究の目標は重複遺伝子を次世代シークエンサで得られた大量配列から同定することである．

2.1　手法の概要

手法の概観を図 2 に示す．本手法の入力は二種類あり，次世代シークエンサで読まれた RNA 配列 (以下リード配列) と近縁種の遺伝子配列である．入力に対し，各リードが既知のどの遺伝子のどの領域に対応するかを知るため，配列アラインメント手法を用いて，リード配列を近縁種の遺伝子配列に対応付ける (図 2(2))．配列アラインメント手法としては，BLAST [1] や BLAT [2] など，既存の手法を用いることができる．アラインメント結果を基に，重複遺伝子の同定を行う (図 2(3)(4))．同定に際し，頻出パターン解析を利用したクラスタリング手法を導入する．

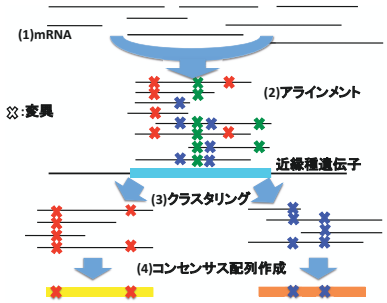

図 2: 重複遺伝子発見の概要

2.2　変異の頻出パターン抽出

本節では頻出パターン抽出と重複遺伝子判定の対応付けをおこない，重複遺伝子判定の手法を導入する．

頻出パターンを発見できるアルゴリズムの一つである Apriori [3] を用いて，頻出する変異を見つける．頻出パターンは，データベースから一定回数以上（最小サポート数以上）現れるアイテム集合である．本研究では，近縁種遺伝子の各塩基をデータベースのトランザクションに対応させ，近縁種遺伝子から変異が認められた場合，変異の位置に相当するトランザクションがリード番号に相当するアイテムを有すると考える．これにより，共通に変異位置を持つリード群を見つけることができる．次に，求まった頻出パターンから，重複遺伝子候補に相当するリード集合を選択する．選択したリード集合の大きさ（リード数）に着目し，手順を分ける．

1. 最長のリード集合が 1 つの場合
   最も長いリード集合が唯一に決まる場合は，そのリード集合をひとつのグループとみなす．その後，次に長いリード集合中で，上記グループとは最も離れているリード集合を探し，再帰的に同様の操作を行う．リード集合間の類似度は，二つのリード集合 R_1, R_2 とし，閾値 r とすると，

   $$D(R_1, R_2) = \frac{r \times \min\{|R_1|, |R_2|\} + \max\{|R_1|, |R_2|\}}{|R_1 \cup R_2|}$$

   で定義する．$|\cdot|$ は要素数である．R_1 と R_2 の間に重複が多いほど，$D(R_1, R_2)$ は大きくなる．

2. 最長のリード集合が 2 つの場合
   $D(R_1, R_2)$ を利用する．閾値 d とすると，$D(R_1, R_2) < d$ の場合に 2 つのグループであるとみなし，終了する．$D(R_1, R_2) \geq d$ の場合，2 つの集合は同一の遺伝子から観測された可能性が高いため，$R_1 \cup R_2$ なる新たな集合を作成し，(1) を実行する．

3. 最長のリード集合が 3 つ以上の場合
   全リード集合間の類似度を，$D(R_1, R_2)$ を用いて計測し，最短距離法によりクラスタを生成する．この操作により，ひとつの集合になった場合には (1) と同様の操作を行い，二つ以上になった場合

には，各集合がそれぞれ異なる遺伝子由来であると考えて終了する．

3　実行結果と考察

ゲノム未知の軟体動物ヒメイカ，ヤリイカ，ホタテガイ，オウムガイから mRNA を採取し，Roche 社 454 Titanium で配列の決定を行った．配列のアラインメントには BLAT を用い，リード配列，近縁種遺伝子配列共に塩基をアミノ酸配列に翻訳したものを使用した．アラインメントは BLAT の出すスコアで E-value が 1e-20 以下かつ，identities の値が 60 以上のものを利用した．また，近縁種遺伝子に対応づけられたリード数が 40 本以下の場合のみ利用している．

頻出パターン抽出の最小サポート c を 0.15，リード集合間の類似度を求める際のパラメータ r を 0.8，閾値 d を 1.0 として実行した．表 1 に BLAT でカサガイの遺伝子に対応したリード数，及び，計算結果から求まった重複遺伝子数の統計を示す．アラインメントされたリード数は，リードの内最低一つのカサガイの遺伝子に対応したリード数，対応遺伝子数は，最低一本のリードが対応している遺伝子数，クラスタ化したリード数は，頻出パターン解析によるリード群作成で最低一つのリード群に含まれるリード数，重複候補の近縁種遺伝子数は，図 1 で A に相当するカサガイの遺伝子数，重複遺伝子候補数は，図 1 で A’ や A” に相当する遺伝子数を示している．

いずれの種においても，カサガイに対応したリードが少なく，本実験系の近縁種がある程度進化的に離れている事が分かる．選んだ近縁種が近くないにも関わらず，オウムガイを除いて残り 3 種では 40 以上の重複遺伝子候補を見つけることができた．

今回の解析で認められた重複遺伝子候補が，互いに異なる機能を有しているかを調査するため，個々のコンセンサス配列を NCBI の BLAST サイトを用いて調査した．データベースには，Non-redundant protein sequences(nr) を用い，全種に対して検索を行った．また，アラインメントには BLASTX を用いた．結果の中から，ヒメイカ，ヤリイカ，ホタテガイからそれぞれ 1, 3, 1 個の重複遺伝子候補を予測した．

表 1: 実行データ及び結果

軟体動物種	ヒメイカ	ヤリイカ	ホタテガイ	オウムガイ
リード数	226,994	262,913	86,832	232,204
アラインメントされたリード数	30,441	5,964	2,594	1,288
対応遺伝子数	879	1,551	652	419
クラスタ化したリード数	753	893	339	23
重複候補の近縁種遺伝子数	92	112	40	4
重複遺伝子候補数	435	467	172	12

4　関連研究

同一種内で配列相同性の高い遺伝子はパラログと呼ばれ，データベースの整備も進むほど [4, 5] 遺伝学には重要な存在である．これらのデータベース作成手法は，ゲノムを決定し，全遺伝子が分かっている種に対し，遺伝子間の相同性を調べ，相同性の高い遺伝子同士をパラログと見なす手法が取られている．しかし，これらの手法はゲノム配列全体あるいはほぼ完全な全遺伝子の配列を要するため，決定までのハードルが高く，モデル生物に限定的な手法である．

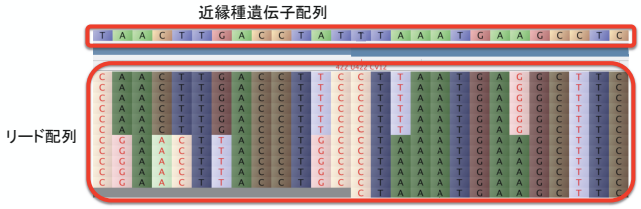

図 3: リードのアラインメント結果

近縁種遺伝子配列	TAACTTGACCTATTTAAATGAAGCCTC
コンセンサス配列1	CAACTTGACCTTCCTTAATGAGGCTTC
コンセンサス配列2	CGAACTTACCTGCCTAAATGAAGCTTC

図 4: 本手法で求められたコンセンサス配列

これに対し，次世代シークエンサによる低コストな配列決定を利用し，モデル生物の近縁種に関しては，部分的な染色体重複を求める手法も利用されている [6]．しかし，この手法もモデル生物の至極近縁種に限定され，多少離れた種でも利用可能である提案手法に優位性がある．

5　まとめ

本研究では，ゲノム未知の種から重複遺伝子を発見するために，相関ルールを用いたクラスタリング手法を導入した．本手法は，次世代シークエンサによって得られた大量配列情報と，進化的に比較的近い種のゲノム・遺伝子配列を用いる事で，変異の共起関係を発見し，それを基に重複遺伝子を予測する手法である．本手法を次世代シークエンサの一つである Roche 社 454 で読んだ軟体動物 4 種の EST 配列に適用する事で，新たな機能を獲得した重複遺伝子を予測する事ができた．

謝辞

本研究を遂行するにあたり，貴重なデータとご助言を数多く賜りましたお茶の水女子大学アカデミックプロダクション特任助教の小倉淳先生，特任研究員の吉田真明さん，チューリヒ大学植物生物学研究所准教授の清水健太郎先生に深く感謝いたします．

参考文献

[1] Stephen F Altschul, Warren Gish, *et al.* Basic local alignment search tool. J. Mol. Biol. Vol. 215, pp. 403-410, 1990.

[2] W. James Kent. BLAT — The BLAST-like alignment tool. Genome Research. Vol. 12, pp. 656-664, 2002.

[3] Rakesh Agrawal and Ramakrishnan Srikant. Fast Algorithms for Mining Association Rules. VLDB, pp. 487-499, 1994.

[4] Magalie Leveugle, Karine Prat, *et al.* ParaDB: a tool for paralogy mapping in vertebrate genomes. Nucleic Acids Research. Vol. 31, No. 1. pp. 63–67. 2003.

[5] Guohui Ding, Yan Sun, *et al.* EPGD: a comprehensive web resource for integrating and displaying eukaryotic paralog/paralogon information. Nucleic Acids Research. Vol. 36, Database Issue, D255-D262. 2008.

[6] Can Alkan, Jeffrey M Kidd, *et al.* Personalized copy number and segnemtal duplication maps using next-generation sequencing. Nature Genetics. Vol. 41, pp. 1061, 2009.